UNIVERSITY *of*
TASMANIA
AUSTRALIA

School of Natural Sciences

Markov Models for the Evolution of Duplicate Genes, and Microsatellites

Tristan Lee Stark
B.Sc. (Hons.)
Supervisors: Dr. Małgorzata O'Reilly & Assoc. Prof. Barbara Holland

February 2018

Submitted in fulfilment of the requirements for the Degree of
Doctor of Philosophy

Declaration

This thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of my knowledge and belief no material previously published or written by another person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright.

This thesis may be made available for loan and limited copying and communication in accordance with the Copyright Act 1968.

Signed:

Tristan Lee Stark

July 3, 2018

Acknowledgements

I can't overstate my gratitude to my two supervisors Dr. Małgorzata O'Reilly and Assoc. Prof. Barbara Holland for their extensive support throughout not only my Ph.D. but my honours before that. You both have been exceedingly generous to me. You have always made time when I have dropped by your offices looking to discuss this or that mathematical or scientific problem, the meticulous edits to drafts you have provided, or matters of bureaucracy. Without both of your consistent input and guidance, this thesis would never have been possible, and I would not have half of the aptitude or enthusiasm for mathematics or science that I have today. Thank you to you both for your care and diligence in all aspects of this mentoring role.

I'd also like to thank everyone from the UTAS theoretical phylogenetics group, all of whom have been engaging colleagues to work with. The group's regular informal meetings, conference organisation and participation, and generally engaging attitude has made my time in the UTAS mathematics department all the more enjoyable. I'd particularly like to thank Dr. Jeremy Sumner for his useful comments on group theory, and the orbit stabiliser theorem.

Thank you also to my wife Catherine, who, particularly in these last few weeks, has taken great care for me personally. Not only have you born the majority of the burden of my survival, and the running of our household, you've also shamed me with your superior LaTeX-skills and helped me to fix my most pernicious typographical errors. You have served as a great mathematical foil over the years, despite your unfortunate lack of interest in the theory of probability. I hope that I serve this function for you as well as you do for me.

Thank you to my close friends and to my family, especially to my mother, for the support that you have provided over the years. To my father, I regret that you won't have the opportunity to see this thesis even more than the one that came before it.

Thanks everyone!

Abstract

Duplicate genes and microsatellites are two key sequences in the study of evolutionary genomics. Gene duplication has been identified as a central process driving functional change in genomes, since it creates functional redundancy in the genome and allows for subsequent mutation to occur in the absence of selective pressure. Microsatellites are rapidly evolving sequences which can be studied over much smaller timescales than most other sequences, and are thus key to the study of population demographics and forensic science.

In this thesis we construct mathematical models for the evolution of duplicate genes, and microsatellites, respectively. We analyse the models in order to make scientific predictions, and derive the following novel results.

We introduce and analyse a modified hazard function, which we use to investigate the preservation of gene duplicates. Further, we construct individual-level models, and present a framework for the extension to population-level models. Also, we construct mappings from mechanistically-motivated intuitive models for gene duplicate evolution, to less intuitive models, which have smaller state spaces and hence are more computationally tractable.

Throughout this analysis, we make scientific predictions based on the properties of the models. We find that the pattern of gene duplicate preservation is more consistent with subfunctionalization than with neofunctionalization. This result is of particular scientific interest, since it is the opposite conclusion of earlier work in the gene duplication literature.

Duplicate genes

Several biological models exist for the evolution of a pair of duplicate genes after a duplication event, and it is believed that gene duplicates can evolve in different ways, according to one process, or a mix of processes. Subfunctionalization is a process under which the two duplicates can be preserved by dividing up the functions of the original gene between them. Here, we find that subfunctionalization is highly consistent with the pattern of gene duplicate preservation, in contrast to previous analysis in the literature.

Another process important to gene duplicate evolution is neofunctionalization, under which both duplicates can be preserved when one copy mutates so as to produce some new beneficial function. Our analysis of neofunctionalization suggests that this process

is not a significant contributor to the preservation of duplicates over the timescales during which regulatory subfunctionalization is resolved. Instead, it is likely that neofunctionalization occurs subsequent to previous subfunctionalization, which acts to preserve copies over the longer time frames required for rare beneficial mutations to have any significant probability of occurring.

Analysis of genomic data using sub- and neofunctionalization models has thus far been relatively coarse-grained, with mathematical treatments usually focusing on the phenomenological features of gene duplicate evolution. In contrast, we develop mechanistically motivated Markov models, and fit directly to duplicate preservation data.

We introduce a modified-cause-specific hazard function to analyse the preservation of gene duplicates. In the context of gene duplication, we refer to this as the pseudogenization rate, owing to the biological interpretation. We analyse the properties of the modified-cause-specific hazard rate in detail, including limit analysis of the general case, and discuss the shape properties of the specific case of the pseudogenization rate.

Further, we extend our model for the evolution of a pair of gene duplicates to model a population of duplicate pairs, by modelling the birth of such pairs as a homogeneous Poisson process. We show that the age distribution of preserved duplicates follows an inhomogenous Poisson distribution, with its rate function depending on the individual-level model. We then fit this distribution to count-data of surviving duplicates in the genomes of four animal species.

Additionally, we extend the individual-level model to a model that includes the process of neofunctionalization, and next, to a model of subfunctionalization for families of gene duplicates. Finally, we map these intuitive models, to less intuitive but more computationally tractable models, and discuss a number of related computational considerations.

Microsatellites

Microsatellites are repetitive regions of DNA where a short motif is repeated many times. Mutations in the number of repeat units occur frequently compared to point mutations and thus provide a useful source of genetic variation for studying recent events. Empirical studies have suggested that the rate of length-changing mutations due to slipped-strand mispairing may depend on the purity of the repeat units, i.e. how well they each match the motif. However, most studies that use microsatellite data are based on models that only track the number of repeat units. In order to address this gap, we introduce a series of models on a two-dimensional state-space

(which are level-dependent quasi-birth-and-death processes) that track the length of the sequence as the level variable, and the number of interruptions (purity) as the phase variable. Our models account for the biological process of point mutation, and its observed effect on the rate of slipped-strand mispairing.

We find that modelling microsatellite purity leads to some complications due to the nature of available data. In terms of the initial model, we discover what constitutes a state-dependent bias in the reporting of repeat sequences by Tandem Repeats Finder (or any similar software used to search whole-genomes for microsatellite sequences). Consequently, we construct a modified model such that all states fall into one of two categories — 'observable states', against which the reporting algorithm is unbiased, and 'unobservable states', which are never reported. We consider two approaches for treating the unobservable states, first to condition on the process being in the observable states, second to treat unobservable states as absorbing. Our initial analysis and underlying biological intuition suggest that transitions from the unobservable to observable states are very rare, and thus we ultimately treat the unobservable states as absorbing.

Additionally, we extend the individual-level model to a population-level model by modelling the birth of microsatellites as a homogeneous Poisson process. We then derive the transient distribution of such model in terms of the individual-level process. This distribution has appropriate relative clock via the inclusion of point mutation. We fit this transient distribution to whole-genome derived sequence data, however we encounter some difficulties in the optimisation owing to the presence of many local optima.

The standard approach for microsatellite models is to make the assumption that the empirical distribution is at equilibrium, and then to fit the stationary distribution to data. The key exception to this is the step-wise mutation model, which predicts infinite growth of the repeat number. Here we fit the above-mentioned transient distribution, and thus do not assume that the empirical distribution is at equilibrium. In contrast to the step-wise mutation model, our model does not predict infinite sequence lengths in the long run.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, we construct and analyse several Markov models for processes related to the evolution of duplicate genes, and microsatellites respectively. The evolution of duplicate genes and microsatellites are both areas of significant interest in evolutionary biology, for reasons discussed in Sections 1.2 and 1.3 respectively. We perform both mathematical, and data-driven analysis using whole-genome derived data for both gene duplicates and microsatellites. We focus on the development of mathematical results, which are put into context through discussion of biological interpretations. With this in mind, we start this chapter by putting this work in context with a brief discussion of the importance of Markov models in evolutionary biology as a whole. We then discuss the details of gene duplicate and microsatellite evolution respectively. Finally, we give an overview of the specific contributions presented in this thesis.

1.1 Markov models in evolutionary biology

The theory of evolutionary biology has become increasingly reliant upon the theory of probability, and in particular the theory of Markov processes, since the 1960s. It is easy to see why the theory of probability is so important to evolution — not only are the mutational events underlying practically all evolutionary processes thought to be inherently random, but the complex interactions between different biochemical structures are often not amenable to direct deterministic analysis, and as such statistical models are often required. Pioneering work by Kimura [64], Jukes and Cantor [59], and Felsenstein [39] have enshrined Markov processes at the heart of models for molecular evolution. Since molecular evolution underlies essentially all of evolutionary biology, this has made the theory of Markov processes central to evolutionary analysis. Beyond just modelling the evolution of proteins or nucleotides, Markov processes have

become a widely used mathematical tool for modelling the evolution of many higher-level genomic processes, including, of particular relevance to this thesis, the evolution of microsatellites.

Despite this, there exists a lot of biological theory which, while amenable, has not been treated with a rigorous Markov-chain based analysis — this is the case for duplicate genes. Also, many of the newer or less well-known results from the mathematical theory are rarely applied in evolutionary biology. Non-stationary models are uncommon, despite their inherent appeal (with the bigger picture of evolution clearly being a non-stationary process), absorbing processes see little application in the evolutionary biology literature. Likewise, the phase-type and matrix exponential distributions are not often applied.

In this thesis, we discuss our research applying rigorous Markov-chain based analysis to two distinct areas of evolutionary biology. The motivation of this research is two-fold — the primary aim is to analyse the systems and contribute to the understanding of evolutionary biology, and the secondary goal is to introduce some overlooked mathematical tools to the evolutionary biology literature.

1.2 Gene duplication

Gene duplication has been identified as a key process driving functional change in many genomes. Several biological models exist for the evolution of a pair of duplicates after a duplication event. Gene duplication was first presented as an important process by Ohno [88], who postulated that the emergence of new functions in genomes was enabled by gene duplication. Gene duplication has since been identified as a common occurrence in sequenced genomes [81], and as an important contributor to genome diversification [55; 78].

It is believed that, after duplication, gene duplicates can evolve according to a range of different biological processes. The central process is *pseudogenization*, where one copy loses its functionality through fixation of deleterious mutations, becoming a so-called *pseudogene* [56]. In the absence of some pressure to preserve both copies, pseudogenization is thought to be the ultimate fate of any duplicate pair.

Ohno [88] proposed the process of neofunctionalization, by which a pair of duplicates could avoid eventual pseudogenization. Ohno [88] claimed that duplication relaxed selective pressures on proteins and enabled mutations to accumulate, and that this could eventually lead one duplicate to gain some new, beneficial functionality, potentially at the expense of existing functionality. In this case, only one copy retains the

ancestral function, and is thus protected from pseudogenization by negative selection. On the other hand, the copy with new functionality could be protected by positive selection, leading to the preservation of both.

Subfunctionalization is a competing hypothesis to explain the preservation of duplicates, which was analysed in a series of papers by Force and Lynch [42; 82; 83]. Subfunctionalization is a process of subdividing functions from the ancestral state between the duplicated gene copies, which allows for both copies of the gene to be preserved by selective pressure without the need to invoke positive selection.

To model the evolution of gene duplicates, sub- and neofunctionalization are (usually separately) treated as processes competing with pseudogenization. Thus, under sub- or neofunctionalization models, the ultimate fate of all duplicates is sub- (respectively neofunctionalization) or pseudogenization.

Force et al. [42] described a process which they referred to as duplication-degeneration-complementation (DDC), which is the essential mechanism by which subfunctionalization is thought to occur. Under the process we have, immediately after some duplication event, two identical genes, each with a fixed number of mutable regulatory regions. Null mutations occurring in the regulatory regions lead to the complementary degeneration of the pair of genes. Functions which are lost in one copy are retained in the other, and vice versa. While either copy on its own would not be sufficient to retain the functionality of the original duplicated gene, together the two copies can do so. As such there is a selective pressure acting to preserve both copies together in the genome. A side-effect of this is that some redundancy will have developed in the regulatory regions which did not undergo null-mutation in either of the copies, and this could lead to further changes by allowing for mutations which in a single copy would be deleterious, but which will not be selected against due to the redundancy created by subfunctionalization. This could allow for other evolutionary processes to search the space of alleles more freely and lead to subsequent neofunctionalization.

Duplicate genes are the subject of Chapters 3 and 4. Our goal was to develop and analyze mechanistically motivated models for the evolution of duplicates after an initial duplication event. Chapter 3 focuses on the analysis of the subfunctionalization process for a pair of duplicates. We build on this work in Chapter 4, where we consider the combined processes of sub- and neofunctionalization for a pair of gene duplicates, and, separately, subfunctionalization for gene families.

Contrary to existing work, we find that neofunctionalization is not a significant contributor to the preservation of duplicate pairs over the timescales during which regulatory subfunctionalization is resolved, and that subfunctionalization predicts the broadly

convex declining pseudogenization hazard rate most often attributed to neofunctionalization.

1.3 Microsatellites

A microsatellite, or simple sequence repeat, is a strand of DNA which repeats a motif of length 1–6 nucleotides [36]. For example, we may have the string of nucleotides ATATATATAT, which is the motif AT repeated 5 times. Microsatellites undergo a mutation process which leads to a change in the number of repeats, at a rate which is orders of magnitude higher than the rate for other forms of mutation, such as point mutation, insertions and deletions [21]. There is some debate as to how many repeats are required for mutations characteristic of microsatellites to occur. Rose and Falush [99] suggest that the threshold is approximately 8 repeats, however more recently Leclercq et al. found that no such threshold exists [73].

Microsatellites are found in vastly greater density than that which would be implied by random allocation of nucleotides [37]. They are found throughout the genome, in coding and non-coding regions and are ubiquitous in prokaryote and eukaryote genomes [118; 133]. Many microsatellites are thought to evolve neutrally, experiencing no selective pressure [37; 124], and polymerase chain reaction techniques lead to a high availability of microsatellite data by allowing for the production of many copies of DNA sequences.

Neutral evolution, together with high levels of polymorphism resulting from frequent mutation, leads to microsatellites being highly favoured as genetic markers (sequences of DNA occurring at a known locus, used to identify an individual or species) [37; 124]. Hence, microsatellites are of interest in a wide array of population genetics and evolutionary inference applications [101].

In order to make inferences using microsatellite data, a biologically realistic model for the time evolution of microsatellites is required [101]. Many theoretical models have failed to explain observed allele frequency distributions [37]. Since most microsatellites seem to evolve neutrally [37; 124], it is reasonable to assume that loci evolve independently and that their repeat numbers are identically distributed. Since we assume that sites do not affect each other, Markov chains are a likely choice for modelling [124], requiring only the additional assumption that future evolution is fully explained (in probability) by the present state. It is important to note that such models are not necessarily valid for microsatellites which experience selective pressure, such as the trinucleotide repeat sequences associated with various disorders [92].

Levinson and Gutman [77] proposed that the biological process of slipped-strand mispairing was the most likely the dominant process underlying the evolution of microsatellites. Slipped-strand mispairing occurs when, during DNA replication, two strands disassociate and a repeat unit in the n^{th} position of the new strand rehybridizes with a complementary repeat unit in some other m^{th} position of the template strand. A loop of unmatched repeats is formed in one strand and the length of the new strand differs from the template by $|n - m|$ repeats. Whether the new strand is longer or shorter than the template, depends on which of the two the loop was formed in. This biological model is widely accepted as the best description of the underlying process (see e.g. [36; 21; 35]).

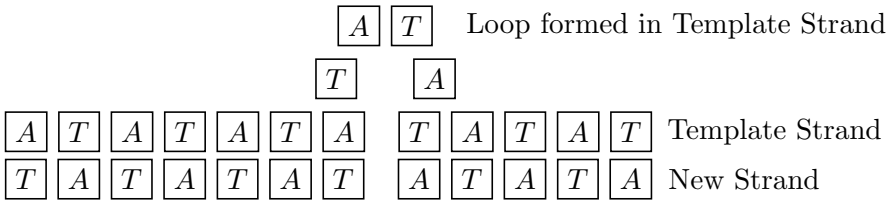

Figure 1.1: Representation of slipped-strand mispairing where the loop is formed in the template strand, shortening the new strand relative to the template.

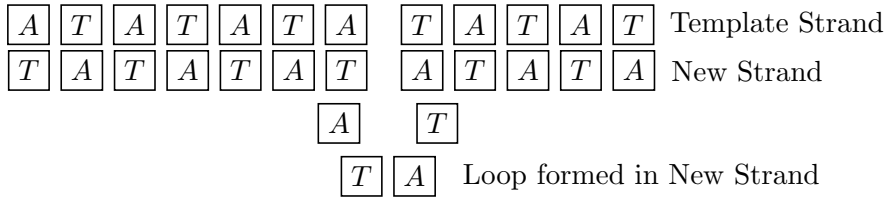

Figure 1.2: Representation of slipped-strand mispairing where the loop is formed in the new strand, increasing its length relative to the template strand.

Two widely used models in the microsatellite literature are the infinite alleles model [64] and the stepwise mutation model [64]. The infinite alleles model assumes that the number of possible alleles is sufficiently large that any mutation necessarily leads to a state not previously existing in the (finite) population, while the stepwise mutation model is a homogeneous birth and death process taking any integer value. Both models are defined in a more general context than that of microsatellite evolution, and are widely used in a range of evolutionary biology contexts.

When microsatellite alleles are identified by their repeat number, as per the slipped strand mispairing process, the restriction of the stepwise mutation model to some subset of the positive integers results in a fairly realistic model for the process. Much work has been done to modify the stepwise mutation model to account for specific

observations of microsatellite loci [37], but the model as introduced by Kimura and Ohta [64] is still widely used for population studies.

1.3.1 Time evolution of microsatellites

There are many factors affecting the way a given microsatellite evolves in time. Below we discuss these factors, and the attempts to account for them in existing microsatellite models. The model which we define in Chapter 5 is built upon the general model due to Wu and Drummond [130], which includes a majority of earlier models as submodels. We restrict their model slightly in view of the findings of other analysis, and then extend it to account for the possibility of interruptions in the repeat sequence.

Size of mutation events

Empirical studies of slipped-strand mispairing have been carried out on a variety of species (see for example [58; 43; 15; 119; 14]), and show that the most common mutations are slippage events in which a single repeat unit is gained or lost. Less frequently, slippage events leading to the gain/loss of multiple repeat units are observed [33]. There is no general agreement on the distribution of mutation sizes, which it is thought to vary between loci [37], but the consensus is that single repeat unit changes are significantly more frequent that changes involving multiple repeat units.

To model large jumps in the repeat number, Di Rienzo [33] proposed a two-phase model. Under this model, given a mutation occurs, it has a probability p of being a one-step mutation, and a probability $(1-p)$ of being a so-called multi-step mutation, in which case the length of the microsatellite is increased or decreased by a number drawn from a specified distribution. Since their work, models have been separated into two major classes, one-phase and two-phase models. While two-phase models are in a strict sense more realistic, since multi-step mutations have been observed empirically, there is little evidence to suggest that they confer any real modelling benefit. In their review of the various models in the literature, Sainudiin et al. [101] found no advantage for two-phase over one-phase models using an Akaike information criterion (AIC) approach. It is likely that the effect of occasional multiple repeat-unit slippage events could be adequately captured by a marginally higher mutation rate without explicitly allowing such events in a model. As such, we restrict our extension of the model due to Wu and Drummond [130] to the one-phase case.

Length dependence of mutation rate

It has been hypothesised that as the length of a microsatellite increases, there is more opportunity for slippage to occur, and hence that the mutation rate increases. This is confirmed experimentally [127; 14; 75; 15], and Ellegren says in his 2004 review [37] "The single most important factor to affect mutation rate that has so far been discovered is microsatellite length."

Kruglyak et al. [68] proposed a model in which a microsatellite of length i goes to length $i+1$ or $i-1$ at a rate $b(i-1)$, where b is a constant. This model provided a good fit for the microsatellites in the yeast genome [67]. Calabrese and Durrett [20] similarly proposed a quadratic model, although Sainudiin [101] found no advantage over the linear model in modelling a Human-Chimpanzee data set. With this in mind, we further restrict the model due to Wu and Drummond [130], upon which we base our model, to allow for only a linear (in repeat number) rate of slipped-strand mispairing.

Contraction vs. expansion — mutational bias

A bias in favour of expansion over contraction for slippage mutations is often observed [24; 60; 34; 95]. However, [47; 127] found a bias in favour of contraction. Xu et al. [131] found that the rate of contractions increased exponentially with repeat length while the rate of expansion remained the same. Other studies have similarly shown a change in directional bias with increasing length [51; 43; 47; 91].

Garza et al. [45] proposed a model using a 'target size' so that mutation was biased to contractions for alleles longer than the target, and towards expansions for shorter alleles. They found that this was sufficient to account for observed allelic variation in humans and chimpanzees. Similar approaches have since been widely adapted by others (e.g. [101; 126]). Wu and Drummond [130] accounted for mutational bias using a logistic function, and we do the same when we extend their model in Chapter 5

We note that in a model biased towards expansion, the length of the microsatellite will inevitably become unrealistically large. Furthermore, if there is a bias towards contraction as well as an increasing rate with length, we would expect to see mostly very short microsatellites. However, if there is a bias in favour of contraction only for microsatellites over a certain length, unending expansion is stymied and we expect to see a more even distribution of lengths.

Point mutations and microsatellite purity

While slipped-strand mispairing accounts for the majority of microsatellite mutations, all DNA is susceptible to point mutation, whereby a single base nucleotide is replaced with another nucleotide. When a point mutation occurs a microsatellite may become interrupted, for example we might have $(AT)_{30}$ become $(AT)_{12}(AC)(AT)_{17}$ [36]. It has been observed [57; 38; 129] that mutation rate varies between pure and interrupted repeats. It stands to reason that an interrupted microsatellite may mutate into a pure one, either by the removal of an impure repeat via slipped-strand mispairing, or by a point mutation correcting the impure repeat. Harr et al. [47] showed that purifying mutations do occur. Since microsatellite repeat number was traditionally measured without identifying all nucleotides at the locus in question, point mutations would often go undetected [21] and thus purity data was relatively more difficult to obtain than repeat length data.

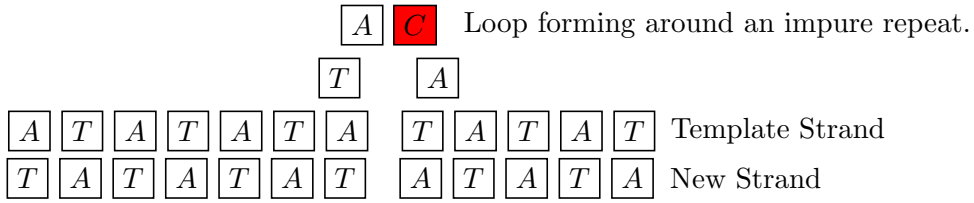

Figure 1.3: Representation of how slipped-strand mispairing might remove an impure repeat.

The essential principle of early DNA sequencing was that sequences of different lengths would distribute at different relative positions when placed in a gel matrix along an electrophoresis plate [102]. Sequences were replicated by polymerase-chain reaction (PCR), and template strands were added to four reaction chambers together with primers designed to bind with the start of the sequence. Subsequently, DNA polymerase, which synthesizes DNA, was added, along with a mixture of the four nucleoside triphosphate molecules necessary to construct the sequence. Finally, to each reaction flask would be added one of four dideoxynucleotide triphosphates, which would cause the sequence to terminate wherever it was bound — sequences in one flask would terminate at A, in another at C, etc. Given sufficiently many template strands, copies of the sequence terminating at every possible location would be produced, which were then run in lanes corresponding to their terminating nucleotide along the electrophoresis plate. Thus, the sequence could be read by reading the order of the lanes in which bands of sequences appear — if the band nearest the positive anode appears in the A lane, the first nucleotide in the sequence is A, and so on. In the case of microsatellite sequences, the process could be simplified such that significantly less amplification of

sequences (in the early days, an extremely costly process) is required.

Due to the repetitive nature of microsatellite sequences, if two appropriate regions flanking the sequence can be identified, primers can be annealed at both ends of the sequence, and the length inferred in a similar manner without identifying the specific nucleotides in the sequence. The length of the sequences can be be inferred directly, along with the size of the repeat unit (since the sequence length would be in integer multiples of the size of the repeat unit, itself in integer multiples of the size of an individual nucleotide). This established the convention in the analysis of microsatellite sequences, which are often still measured in a similar manner, but with the advent of modern DNA sequencing techniques, full sequencing is the norm, and plenty of sequence data is now available.

Attempts to reconcile point mutation and slipped-strand mispairing have led to counting schemes that eliminate information about the purity of a sequence in order to maintain both a one dimensional state space in the models, and the convention established by history of DNA sequencing. For example, Calabrese [20] counts only uninterrupted repeats, Sibly [106] counts only the left half of an interrupted repeat and Bell and Jurka [9] tracked either side of an interruption as individual, pure microsatellites. Given the effect an impurity has on mutation rate, we expect that accounting for varying levels of purity may improve the models. To that end, in Chapter 5, we extend the model due to Wu and Drummond [130] (with the restrictions mentioned above) to explicitly account for the number of interruptions in the repeat sequence.

Aside from those mentioned above, a variety of other factors can influence mutation rate, including sex, the particular repeat motif, and the locus at which it occurs [15; 14; 34]. However, aside from sex, these factors are constant over the life-cycle of a microsatellite. As such, their effects can be accounted for in the choice of parameters of the model, rather than needing to be explicitly accounted for. For example, to account for different mutation rates for varying repeat motifs, a model can be fit to data corresponding to the different motifs.

Microsatellites are the topic of Chapter 5. Our goal in examining the evolution of microsatellites was to develop and analyse a model which would take into account the important features already considered in the literature, as well as incorporating point mutation and sequence interruption. Whole-genome derived microsatellite data inclusive of interruptions poses some specific challenges which we attempt to account for in the model development. Ultimately, we were not able to make any strong conclusions due to the quality of the data. Nonetheless, some interesting mathematical results arise, and we are able to make some scientific predictions in spite of the quality

of the data.

1.4 Thesis overview

In Chapter 2 we provide a summary of some important, well-known results from the theory of Markov processes and statistics, together with some intuitions regarding the interpretation of certain statistical measures. The first part of this summary is used to establish notational conventions and nomenclature for the subsequent chapters, while the later parts highlight some points of particular interest.

In Chapters 3 and 4 we examine gene duplication. The evolution of gene duplicates is a part of the biological theory which has seen relatively little application of Markov processes. We define Markov models for the evolution of the descendents of some gene which has been duplicated by some (unspecified) process. After duplication, the evolution of the resulting copies of the gene is eventually resolved according to one of several competing and/or complementary biological models. Central to this evolution is the process of pseudogenization, whereby copies of the gene can be nonfunctionalized, and essentially lost to the genome (in the sense that it is no longer functional, and subject to potential deletion, or further degradation). Other processes compete with pseudogenization to fix the copies under selective pressure, ensuring their preservation in the genome. We model pseudogenization alongside two different biological models for the preservation of gene duplicates.

In Chapter 3 we consider the evolution of a pair of gene duplicates under the biological process of subfunctionalization. The duplication-degeneration-complementation (DDC) mechanism, which is detailed by Force et al. [42], describes in precise terms how the mechanics of evolution by subfunctionalization occur for genes after a duplication event. Early work by Force [42], and Force and Lynch [82; 83] considered subfunctionalization in terms of competing Poisson processes. Subsequent mathematical modelling of the process has been relatively ad hoc [54], employing unjustified approximation where an exact Markov-chain based analysis is possible. We have performed such an analysis in our recent paper [109], and we discuss this work and some related unpublished results in Chapter 3. As part of this analysis we introduce a modified-cause-specific hazard rate, and we extend our model for a pair of duplicates to a model for a population of such pairs in order to fit to whole-genome derived count-data.

In our recent publication [109], together with Dr. Małgorzata O'Reilly, Assoc. Prof. Barbara Holland and Assoc. Prof. David Liberles, we modelled a population of dupli-

cate genes undergoing subfunctionalization by combining an absorbing Markov process describing the evolution of a pair of duplicates with a Poisson duplication process. My contribution to this work included the conception of the model (together with Dr. O'Reilly and Assoc. Prof. Holland), the derivation of a majority of results, coding and data analysis, and writing the majority of the drafts and the final manuscript. Dr. O'Reilly derived the results for 'Probabilities corresponding to i-th mutational events' (Section A in Additional file 1 of [109]) as well as 'Other measures of interest' (Section B.4 in Additional file 1 of [109]), and contributed to the derivation of various other results, wrote part of the initial draft and edited subsequent drafts. Assoc. Prof. Holland contributed to the derivation of various results, edited drafts, and provided biological insights. Assoc. Prof. Liberles edited drafts and provided key biological insights and interpretation of results in the context of gene duplication. These contributions are included in Chapter 3 of this thesis.

In Chapter 4, we introduce some further models related to gene duplication. In particular, we extend the model from Chapter 3 to model gene duplicates evolving under the combined processes of sub- and neofunctionalization. We analyse this model, including extension to the population-level, and fitting to the same dataset considered in Chapter 3. Contrary to previous analysis in the literature, we conclude that subfunctionalization is the dominant mode of preservation of gene duplicates, with neofunctionalization only occurring with any significant probability after earlier subfunctionalization. Separately, we extend the model from Chapter 3 to model the evolution of a family of gene duplicates evolving under subfunctionalization. Mechanistically modelling gene families is significantly more complex than pairs of duplicates, and we introduce a procedure for model development which avoids the need to individually consider the many possible transitions the process can undergo. We explicitly consider the problem of evaluating the state space corresponding to gene families of fixed or dynamic size, and outline a procedure to efficiently compute the associated generator matrix.

In Chapter 5, we propose several models for the evolution of individual microsatellites which treat the level of interruption of the repeat sequence explicitly. Existing models do not account for interruptions in the repeat sequence, however it is well demonstrated that the dynamics of microsatellite evolution vary between pure and interrupted repeat sequences [57; 38; 129]. The model which we ultimately employ for data analysis is an absorbing level-dependent quasi-birth-and-death process, in which the level tracks the length of the sequence, and the phase tracks the number of interruptions. We define an absorbing boundary in terms of the relative level of interruption in the repeat sequence, which is matched to the constraints of available data.

Existing models are designed to have unique stationary distribution, which is fit to empirical data assumed to be at equilibrium [21]. The prevailing biological theory suggests a different picture, with microsatellites thought to undergo finite life-cycles [18; 85; 96; 114; 19]. We extend the individual-level model to a population of microsatellites, and derive a transient distribution with appropriate relative clock to fit to empirical data which is not assumed to be at equilibrium.

However, we find that estimating globally optimal parameters (i.e. maximum likelihood estimates) is not achievable with our optimisation routine. Nonetheless, the concentration of parameter estimates around values associated with an extreme slowdown of mutation for interrupted sequences are indicative that the constraints of the data collection were overly permissive of highly interrupted sequences. We conclude that the dataset is likely to be so polluted with non-microsatellite sequences that even ignoring optimisation issues, estimates derived from the dataset are not likely to be representative of genuine impure microsatellites. Further work is needed to find a reliable optimization routine for this problem, and to clean the dataset to ensure that observations are highly likely to be microsatellite sequences.

Chapter 2

Mathematical Prerequisites

In this chapter, we provide some statements of existing results which will be relied upon throughout the following chapters of the thesis. The main purpose of this chapter is to establish the conventions which we will follow in subsequent chapters, and to provide in-text references for important results. Proofs are not provided for the results stated in this chapter (references to proofs are).

In Section 2.1 we state a selection of key results from the theory of Markov processes. Markov processes are central to this thesis, and readers are assumed to be familiar with the theory. Sections 2.1.1 and 2.1.2 serve primarily to establish notational and nomenclatural conventions which will be followed in the subsequent chapters. As such, throughout Sections 2.1.1 and 2.1.2 statements are given with little exposition. We give some more details in Section 2.1.3 where we discuss absorbing continuous-time Markov chains, which are particularly important to this thesis. For an introduction to the theory of Markov processes, we suggest Ross's 'Introduction to Probability Models' [100], Karlin and Taylor's 'An Introduction to Stochastic Modeling' [115], or Karlkarni's 'Modeling and Analysis of Stochastic Systems' [70]. More advanced topics relevant to this thesis are covered by Neuts's 'Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach' [87] and Latouche and Ramaswami's 'Introduction to Matrix Analytic Methods in Stochastic Models' [72]. Further, the review of quasi-stationary distributions provided by van Doorn and Pollett [121] is highly expository for that topic.

In Section 2.2 we discuss the statistical notion of likelihood, and some approaches to the statistical problem of model selection. We employ some results from the theory of model selection at various points in the subsequent chapters, but it is not central to the thesis in the same sense as Markov processes. We assume that readers have a

basic familiarity with statistical theory, but we provide some exposition in terms of intutions regarding particular results which we rely upon. For a detailed discussion of likelihood we suggest Rohatgi's 'An Introduction to Probability and Statistics' [98], and for model selection we suggest Burnham and Anderson's 'Model Selection and Multimodel Inference' [17].

2.1 Markov processes

The theory of Markov processes provides a set of powerful probabilistic tools with which to model real world systems, and is ubiquitous in the context of evolutionary biology. Throughout this thesis we will be making frequent use of the theory, and what follows is a summary of a selection of important results.

2.1.1 Discrete-time Markov chains

We are principally interested in the application of continuous-time Markov chains in this thesis. However, Discrete- and Continuous-time Markov chains are closely related, and in keeping with Ross [100], we start by introducing the discrete-time analogue.

Definition 1 (Discrete-time stochastic process)**.**

A discrete-time stochastic process is a sequence $X = \{X_n : n \geqslant 0\}$*, where* X_n *is a random variable for each* $n \in \mathbb{N} = \{0, 1, 2, \ldots\}$*. If* $X_n = i$ *we say that* X *is in state* i *at time* n*.*

The set $\mathcal{S}$ *of values taken by* X_n *is called the state space. If* $\mathcal{S}$ *is discrete we say that* X *is discrete valued.*

All of the Markov chains discussed throughout this thesis will be discrete valued.

Definition 2 (Discrete-time Markov chain)**.**

Let $\{X_n : n \in \mathbb{N}\}$ *be a discrete valued stochastic process in discrete time, with state space* $\mathcal{S}$*. We say that such a process is a discrete-time Markov chain (DTMC) if it has the property that for any states* $i_1, ..., i_{n-1}, i, j \in \mathcal{S}$ *and any* $n \in \mathbb{N}$*, we have*

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1) = P(X_{n+1} = j \mid X_n = i). \quad (2.1)$$

Equation (2.1) is called the *Markov property* for discrete-time processes.

Definition 3 (Time homogeneity)**.**

A time-homogeneous Markov chain is a DTMC for which transition probabilities do not depend on the time n, so we have, for all $i, j \in \mathcal{S}, n \in \mathbb{N}$

$$P(X_{n+1} = j \mid X_n = i) = P(X_1 = j \mid X_0 = i).$$

Definition 4 (One-step transition probability matrix)**.**

In the time-homogeneous case we let $P_{ij} = P(X_{n+1} = j \mid X_n = i)$ and define one-step transition probability matrix $\boldsymbol{P} = [P_{ij}]_{i,j\in\mathcal{S}}$ so that

$$\boldsymbol{P} = \begin{bmatrix} P_{00} & P_{01} & \dots & P_{0j} & \dots \\ P_{10} & P_{11} & \dots & P_{1j} & \dots \\ \vdots & \vdots & & \vdots & \\ P_{i0} & P_{i1} & \dots & P_{ij} & \dots \\ \vdots & \vdots & & \vdots & \end{bmatrix}.$$

Definition 5 (n-step transition probability matrix)**.**

For any $i, j \in \mathcal{S}$ and $n = 0, 1, 2, \dots$, define

$$P_{ij}^{(n)} = P(X_{n+k} = j \mid X_k = i),$$

interpreted as the probability that a process in state i will be in state j after n transitions (in n steps). Further, define n-step transition probability matrix $\boldsymbol{P}^{(n)} = [P_{ij}^{(n)}]_{i,j\in\mathcal{S}}$.

The following theorem and corollary are demonstrated in Section 4.2 of Ross [100].

Theorem 1 (Chapman–Kolmogorov equations)**.**

For all $n, m \in \mathbb{N}$ and $i, j \in \mathcal{S}$, we have

$$P_{ij}^{(n+m)} = \sum_{k\in\mathcal{S}} P_{ik}^{(n)} P_{kj}^{(m)},$$

that is,

$$\boldsymbol{P}^{(n+m)} = \boldsymbol{P}^{(n)} \boldsymbol{P}^{(m)}.$$

Corollary 1.

For all $n = 1, 2, \dots$, we have

$$\boldsymbol{P}^{(n)} = \boldsymbol{P}^n. \tag{2.2}$$

Definition 6 (Accessibility)**.**

If $P_{ij}^n > 0$, for some $n \geqslant 0$ we say that state j is accessible from state i, and write $i \to j$.

Definition 7 (Communication)**.**

If states i and j are accessible from each other, they are said to communicate and we denote the relation defined by communication $i \leftrightarrow j$.

The following theorem is demonstrated in Section 4.3 of Ross [100].

Theorem 2 (Communication is an equivalence relation)**.**

The relation defined by $\leftrightarrow$ is an equivalence relation, and hence partitions the state space.

Definition 8 (Communicating class)**.**

We refer to the equivalence classes of $\leftrightarrow$ as communicating classes.

Definition 9 (Irreducible Markov chain)**.**

We say that a Markov chain is irreducible, if every state in $\mathcal{S}$ communicates with all other states in $\mathcal{S}$. That is, $i \leftrightarrow j$ for all $i, j \in \mathcal{S}$.

Definition 10 (First return time)**.**

We define τ_i be the time to first return to state i, given the process started there, i.e.

$$\tau_i = \begin{cases} \infty & \text{if } X_n \neq i, \forall\, n \geqslant 1 \\ min\{n \geqslant 1 : X_n = i \mid X_0 = i\} & \text{otherwise.} \end{cases} \tag{2.3}$$

Definition 11 (Probability of return)**.**

The probability of ever returning to state i is defined as $f_i = P(\tau_i < \infty)$.

Definition 12 (Reccurence and Transience)**.**

We call a state i recurrent if $f_i = 1$ and transient if $f_i < 1$. We call a Markov chain recurrent if all its states are recurrent, and transient otherwise.

The following proposition is demonstrated in Section 4.3 of Ross [100].

Proposition 1.

Let C be a communicating class. Then, if $i \in C$ is recurrent, so is j for any $j \in C$.

For an irreducible Markov chain, either all states are recurrent, or all are transient. Further, a finite-state irreducible Markov chain is necessarily recurrent, since it must make infinitely many transitions among finitely-many states, at least one (and hence all) must be recurrent.

Definition 13 (Mean recurrence time)**.**

We define the mean recurrence time to state i by $M_i = E(\tau_i)$.

Definition 14 (Positive and null recurrence)**.**

We call a recurrent state i positive recurrent if M_i is finite, and null recurrent otherwise. We call say a DTMC is positive recurrent if all of its states are positive recurrent.

Definition 15 (Periodicity)**.**

If d is the largest integer such that $P_{ii}^{(n)} = 0$ whenever n is not divisible by d, we say state i has period d. If $d = 1$ we say i is aperiodic, and periodic otherwise.

Definition 16 (Ergodicity)**.**

A state i is called ergodic if it is positive recurrent and aperiodic. A DTMC is called ergodic if all of its states are ergodic.

Definition 17 (Stationary distribution)**.**

We refer to a vector $\underline{\pi}$ such that

$$\underline{\pi}\mathbf{P} = \underline{\pi}, \tag{2.4}$$

and

$$\sum_{j \in \mathcal{S}} \pi_j = 1, \tag{2.5}$$

as a stationary distribution.

The following proposition appears as Proposition 4.4 in Ross [100].

Proposition 2.

A Markov chain $\{X_n\}$ has a stationary distribution if and only if it is positive recurrent, and when it exists it is given by

$$\pi_j = \frac{1}{M_j}. \tag{2.6}$$

Definition 18 (Limiting distribution)**.**

We refer to $\underline{\pi}^* = [\pi_j^*]$ as the limiting distribution where

$$\pi_j^* = \lim_{n \to \infty} \frac{1}{n} \sum_{m=0}^{n} P_{ij}^m, \tag{2.7}$$

given that the limit below exists and is independent of i for all j.

When the limiting distribution exists, π_j^* is equal to the long-run proportion of time the process spends in state j.

The next theorem follows from Theorem 4.1 in Ross [100].

Theorem 3.

For an irreducible ergodic DTMC the limiting distribution exists, and is equal to the unique stationary distribution.

2.1.2 Continuous-time Markov chains

Continuous-time Markov chains are continuous-time analogous to discrete-time Markov chains. Continuous-time Markov chains are of principle interest in this thesis, and we provide a little more exposition of the related theory.

Definition 19 (Continuous-time Markov chain)**.**

We call a stochastic process $\{X(t) : t \geqslant 0\}$ with discrete state space $\mathcal{S}$ a continuous-time Markov chain (CTMC) if for all $t \geqslant 0, s > u \geqslant 0, i, j, x(u) \in \mathcal{S}$,

$$P(X(t+s) = j \mid X(s) = i, X(u) = x(u)) = P(X(s+t) = j \mid X(s) = i). \tag{2.8}$$

Equation (2.8) is the Markov property for continuous-time processes. It is analogous to the Markov property for the discrete-time case in that it defines a system whose future evolution depends on its history only through the present state. In continuous-time the Markov property specifies that transition probabilities do not depend on the time spent in a particular state.

Definition 20 (Time homogeneity)**.**

We call a CTMC time-homogeneous if for all $t, s \geqslant 0, i, j \in \mathcal{S}$, we have

$$P(X(t+s) = j \mid X(s) = i) = P(X(t) = j \mid X(0) = i). \tag{2.9}$$

We will be considering time-homogeneous Markov chains throughout this thesis.

Definition 21 (Transition matrix)**.**

In the case of a time-homogeneous CTMC, we define the transition matrix

$$\mathbf{P}(t) = [P_{ij}(t)]_{i,j \in \mathcal{S}},$$

where for all $i, j \in \mathcal{S}$, $t \geqslant 0$,

$$P_{ij}(t) = P(X(t) = j \mid X(0) = i). \tag{2.10}$$

Definition 22 (The generator matrix)**.**

We define generator matrix $\mathbf{Q} = [q_{ij}]$ *such that*

$$\mathbf{Q} = \frac{d}{dt}\mathbf{P}(t)\Big|_{t=0} = \mathbf{P}'(0). \tag{2.11}$$

Definition 23 (Holding time)**.**

Assuming the process is in state i *at time* 0*, we define*

$$H_i = \inf\{t > 0 : X(t) \neq i\},$$

referred to as the holding time in state i*.* H_i *is a strictly positive continuous random variable.*

The following proposition is demonstrated in Sections 5.2.2 and 6.2 of Ross [100].

Proposition 3.

For all $i \in \mathcal{S}$*,* $H_i \sim \text{Exp}(-q_{ii})$*. That is, the holding time for a CTMC is necessarily exponentially distributed with parameter* $-q_{ii}$*.*

The following three theorems are demonstrated by Lemma 6.3, Theorem 6.1, and Theorem 6.2 of Ross [100] respectively.

Theorem 4 (Chapman–Kolmogorov equations)**.**

For all $t \geqslant 0, s \geqslant 0$*,*

$$\boldsymbol{P}(t+s) = \boldsymbol{P}(s)\boldsymbol{P}(t), \tag{2.12}$$

or equivalently, for all $t \geqslant 0, s \geqslant 0, i, j \in \mathcal{S}$*,*

$$P_{ij}(t+s) = \sum_{k\in\mathcal{S}} P_{ik}(s)P_{kj}(t). \tag{2.13}$$

Theorem 5 (Kolmogorov backward equations)**.**

For all $i, j \in \mathcal{S}$*,* $t \geqslant 0$

$$P'_{ij}(t) = \sum_k q_{ik}P_{kj}(t), \tag{2.14}$$

or equivalently,

$$\mathbf{P}'(t) = \mathbf{Q}\mathbf{P}(t). \tag{2.15}$$

Theorem 6 (Kolmogorov forward equations)**.**

Under certain regularity conditions (see remark below) we have, for all $i, j \in \mathcal{S}$*,* $t \geqslant 0$*,*

$$P'_{ij}(t) = \sum_k q_{kj}P_{ik}(t), \tag{2.16}$$

or equivalently

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{Q}. \tag{2.17}$$

Remark 1.

The regularity conditions of Theorem 6 are satisfied whenever the process undergoes at most finitely many transitions in a finite time. This is trivially satisfied when the state space is finite, and will be satisfied for all of the models discussed in this thesis.

Section 6.9 of Ross [100] establishes the following proposition.

Proposition 4.

The solution to the Kolmogorov backward and forward equations is $\boldsymbol{P}(t) = e^{\boldsymbol{Q}t}$.

Definition 24 (Accessibility)**.**

If $P_{ij}(t) > 0$ *for some* $t \geqslant 0$*, we say that* j *is accessible from* i*.*

Definition 25 (Communication)**.**

If i *and* j *are accessible from each other, they are said to communicate. We denote the relation thus defined by* $i \leftrightarrow j$*.*

As in the discrete-time case, $\leftrightarrow$ is an equivalence relation.

Definition 26 (Communicating class)**.**

We refer to the equivalence classes of $\leftrightarrow$ *as communicating classes.*

Definition 27 (Irreducible continuous-time Markov chain)**.**

We say that a CTMC is irreducible if $i \leftrightarrow j$ *for all* $i, j \in \mathcal{S}$*.*

Definition 28 (Embedded chain)**.**

Let t_n *denote the time at which the* n^{th} *transition from some state* i *to some other state* j *occurs and let*

$$X_n = \begin{cases} X(0) & \text{for } n = 0 \\ \lim\limits_{t \to t_n^+} X(t) & \text{for } n \geqslant 1. \end{cases} \tag{2.18}$$

Then $\{X_n : n \geqslant 0\}$ *is a DTMC tracking the state changes of the CTMC. We refer to* $\{X_n : n \geqslant 0\}$ *as the embedded chain.*

Notice that X_n is the value taken by $X(t)$ immediately after the n^{th} change of state.

The following proposition is established in Chapter 6 (immediately preceding Theorem 6.8) of Kulkarni [70].

Proposition 5.

The one step transition matrix $\mathbf{P} = [P_{ij}]$ *of the embedded chain is given by*

$$P_{ij} = \begin{cases} \dfrac{q_{ij}}{-q_{ii}} & \text{if } q_{ii} \neq 0, i \neq j \\ 0 & \text{if } q_{ii} \neq 0, i = j \\ 0 & \text{if } q_{ii} = 0, i \neq j \\ 1 & \text{if } q_{ii} = 0, i = j. \end{cases} \tag{2.19}$$

The following proposition follows from Theorem 6.8 of Kulkarni [70].

Proposition 6.

A CTMC is irreducible if and only if its embedded chain is irreducible.

Definition 29 (Time to return)**.**

Define

$$\tau_i = \inf\{t > s : X(t) = i \mid X(0) = i, X(s) \neq i\}, \tag{2.20}$$

interpreted as the time taken for the process to return to state i *given that it started there.*

Definition 30 (Recurrence and Transience)**.**

We call a state recurrent if $P(\tau_i < \infty) = 1$*, and transient otherwise.*

Definition 31 (Positive-recurrence and null-recurrence)**.**

If a state i *is recurrent, and* $E(\tau_i) < \infty$*, we say the state* i *is positive-recurrent, and if* $E(\tau_i = \infty)$ *we say* i *is null-recurrent.*

Theorem 6.9 of Kulkarni [70] proves the following.

Proposition 7.

A state i *of a CTMC is recurrent (transient) if and only if it is recurrent (transient) in the embedded chain.*

The same does not apply for positive- and null-recurrence.

Theorems 6.9 and 6.10 of Kulkarni [70] prove that, under the regularity conditions mentioned in Remark 1, transience, recurrence, positive-recurrence and null-recurrence are all class properties (in the sense of communication classes). The next theorem follows from this result.

Theorem 7.

For a regular irreducible CTMC all states are together transient, positive-recurrent, or null-recurrent.

As in the discrete-time case we call a CTMC positive-recurrent, null-recurrent or transient if it is irreducible and all states are such.

Definition 32 (Stationary distribution)**.**

We call a vector $\underline{\pi} = [\pi_j]$ *a stationary distribution if, for all* $t \geqslant 0$*,*

$$\underline{\pi}\mathbf{P}(t) = \underline{\pi}, \tag{2.21}$$

and,

$$\sum_{j \in \mathcal{S}} \pi_j = 1. \tag{2.22}$$

The following corollary is established in Section 6.5 of Ross [100].

Corollary 2.

Any stationary distribution satisfies

$$\underline{\pi}\mathbf{Q} = \underline{0}, \tag{2.23}$$

or equivalently

$$-q_{jj}\pi_j = \sum_{\substack{k \in \mathcal{S} \\ k \neq j}} \pi_k q_{kj}, \tag{2.24}$$

where $\underline{0}$ *represents a vector of zeros of appropriate size.*

We call the system of equations defined by Equation (2.23) the *balance equations.*

Remark 2.

Throughout, we will use $\underline{0}, \underline{1}$ *to denote vectors of zeros and ones of appropriate size respectively. Likewise, we use* $\mathbf{0}$ *and* $\mathbf{1}$ *to denote matrices full of zeroes and ones of appropriate size respectively. We use* $\underline{e}_i$ *to represent a vector with a one in the* i^{th} *entry and zeros elsewhere.*

Definition 33 (Limiting distribution)**.**

Assuming the limits exist and are independent of i*, we define limiting distribution* $\underline{\pi}^* = [\pi_j^*]$*, where the limiting probabilities* π_j^* *for each* $j \in \mathcal{S}$ *are given by*

$$\pi_j^* = \lim_{t \to \infty} P_{ij}(t). \tag{2.25}$$

Definition 34 (Ergodicity)**.**

We say that a CTMC is ergodic when the limiting distribution $\underline{\pi}^*$ exists.

The following proposition is established in Section 6.5 of Ross [100].

Proposition 8.

Given an irreducible, positive recurrent CTMC, the limiting distribution exists, and is equal to the unique stationary distribution.

2.1.3 Absorbing CTMCs

Definition 35 (Closed set)**.**

For any $\mathcal{J} \subseteq \mathcal{S}$, if $P_{ij}(t) \neq 0$ for some $t \geqslant 0$ implies $j \in \mathcal{J}$ for all $i \in \mathcal{J}$ then we say the set $\mathcal{J}$ is closed.

Definition 36 (Absorbing state)**.**

If for all $j \neq i$, $i, j \in \mathcal{S}$, $P_{ij}(t) = 0$ for all $t \geqslant 0$ then we say that state i is absorbing.

Definition 37 (Absorbing set)**.**

We call the collection $\mathcal{A}$ of all absorbing states of $\mathcal{S}$ the absorbing set, and $\mathcal{A}$ is necessarily closed.

Definition 38 (Absorbing CTMC)**.**

We call a CTMC $\{X(t) : t \geqslant 0\}$ absorbing if every state is either absorbing or transient.

Not all CTMCs with absorbing states fit our definition of an 'absorbing CTMC'. The results which follow do not necessarily hold for such processes.

Definition 39 (Transient set)**.**

For an absorbing CTMC with state space $\mathcal{S}$ and absorbing set $\mathcal{A}$ we call the set of transient states $\mathcal{S}^* = S \backslash \mathcal{A}$ the transient set. We say that the transient set is irreducible if it is a communicating class.

For the following discussion, suppose that an absorbing CTMC $\{X(t) : t \geqslant 0\}$ has state space $\mathcal{S}$, absorbing set $\mathcal{A}$, transient set $\mathcal{S}^*$ and generator matrix $\mathbf{Q} = [q_{ij}]$. Denote the j^{th} (arbitrarily ordered) absorbing state by a_j, and let $\underline{v}_j = [v_{ij}]_{i \in \mathcal{S}^*}$ be a vector such that $v_{ij} = q_{ia_j}$ for all $i \in \mathcal{S}^*$ for each $a_j \in \mathcal{A}$. Further, let $\mathbf{V} = [v_{ij}]_{i \in \mathcal{S}, a_j \in \mathcal{A}}$ be a matrix with the j^{th} column equal to $\underline{v}_j$.

Definition 40 (Subgenerator matrix)**.**

We define the subgenerator *matrix by* $\mathbf{Q}^* = [q_{ij}]_{i,j\in\mathcal{S}^*}$.

The canonical form of the generator matrix $\mathbf{Q}$ is given by the block matrix form

$$\mathbf{Q} = \left[\begin{array}{c|c} \mathbf{Q}^* & \mathbf{V} \\ \hline \mathbf{O} & \mathbf{O} \end{array}\right]. \tag{2.26}$$

In the case of a single absorbing state, $\mathbf{V}$ is a vector, and we write $\mathbf{V} = \underline{v}$ so that the canonical form of the generator is given by

$$\mathbf{Q} = \left[\begin{array}{c|c} \mathbf{Q}^* & \underline{v} \\ \hline \underline{0} & 0 \end{array}\right]. \tag{2.27}$$

The following proposition and theorem are proved by Theorem 2.4.3 in Latouche and Ramaswami [72].

Proposition 9.

The subgenerator matrix $\mathbf{Q}^*$ *of an absorbing CTMC is invertible.*

Theorem 8.

Given an absorbing CTMC, $\lim_{t\to\infty} P(X(t) = i) = 0$ *for all* $i \in \mathcal{S}^*$. *That is, absorption into some* $j \in \mathcal{A}$ *occurs eventually with probability equal to* 1.

Definition 41 (Time-to-absorption)**.**

We call random variable $T = \min\{t : X(t) = i, i \in \mathcal{A}\}$ *the time-to-absorption.*

Definition 42 (Phase-type distribution)**.**

Consider an absorbing CTMC $\{X(t) : t \geqslant 0\}$ *and initial distribution* $\underline{\alpha} = [\alpha_i]$, *where* $\alpha_i = P(X(0) = i)$ *for all* $i \in \mathcal{S}$. *The phase-type distribution is the distribution of time-to-absorption of such a process, and is parameterized by the subgenerator matrix* $\mathbf{Q}^*$ *together with initial distribution* $\underline{\alpha}$*; we write* $T \sim PH(\mathbf{Q}^*, \underline{\alpha})$ *to denote such distribution.*

Most often, absorbing CTMCs are considered with only one absorbing state, and the phase-type distribution is usually defined as such. Throughout this thesis we will be particularly interested in absorbing CTMCs with multiple absorbing states. Therefore, we give a corresponding treatment of the phase-type distribution.

A proof of the following theorem follows from a very slight modification (to account for multiple absorbing states) of Theorem 2.4.1 in Latouche and Ramaswami [72].

Theorem 9.

The phase-type distribution $PH(\mathbf{Q}^*, \underline{\alpha})$ has cumulative distribution function

$$F(t) = P(T < t) = 1 - \underline{\alpha} e^{\mathbf{Q}^* t} \underline{1}. \tag{2.28}$$

Further, $PH(\mathbf{Q}^*, \underline{\alpha})$ has probability density function

$$\begin{aligned} f(t) = F'(t) &= -\underline{\alpha} e^{\mathbf{Q}^* t} Q^* \underline{1} \\ &= \underline{\alpha} e^{\mathbf{Q}^* t} \mathbf{V} \underline{1}. \end{aligned} \tag{2.29}$$

The two alternative forms are due to the property $\mathbf{Q}\underline{1} = \underline{0}$ of the generator of the Markov chain, which gives

$$\begin{aligned} \mathbf{Q}^* \underline{1} + \mathbf{V}\underline{1} &= \underline{0}, \\ \mathbf{Q}^* \underline{1} &= -\mathbf{V}\underline{1}. \end{aligned} \tag{2.30}$$

In the case of a single absorbing state $\mathbf{V}\underline{1}$ is replaced throughout by $\underline{v}$.

Definition 43 (Hazard function)**.**

The hazard function $\lambda_i(t)$ given that the process starts in state $i \in \mathcal{S}^*$ is defined for all $t \geqslant 0$ as,

$$\lambda_i(t) = \lim_{h \to 0^+} \frac{P(t < T < t + h \mid T > t, X(0) = i)}{h} = \frac{f_i(t)}{1 - F_i(t)}, \tag{2.31}$$

where $f_i(t)$ is the probability density of absorption occurring at time t given that the process starts in state i, and $F_i(t)$ is the corresponding cumulative distribution function.

The hazard function can be interpreted as the conditional (on not having been absorbed before time t) expected exponential rate of absorption at time t.

Definition 44 (Survival function)**.**

The survival function $S_i(t)$ given that the process starts in state i is defined for all $t \geqslant 0$ as,

$$S_i(t) = 1 - F_i(t). \tag{2.32}$$

The survival function can be interpreted as the probability that the process has not been absorbed by time t.

The next theorem follows from Theorem 9 (Theorem 2.4.1 in [72]), and the fact that $T \sim PH(\mathbf{Q}^*, \underline{e}_i)$.

Theorem 10.

For all $i \in \mathcal{S}^*$ and all $t \geqslant 0$, we have

$$\lambda_i(t) = \frac{\underline{e}_i e^{\boldsymbol{Q}^* t} \mathbf{V} \underline{1}}{\underline{e}_i e^{\boldsymbol{Q}^* t} \underline{1}}, \tag{2.33}$$

and

$$S_i(t) = \underline{e}_i e^{\boldsymbol{Q}^* t} \underline{1}. \tag{2.34}$$

Definition 45 (Time-to-absorption into j, $\mathcal{J}$)**.**

For each state $j \in \mathcal{A}$, we define the time-to-absorption into j by $T_j = \min\{t : X(t) = j\}$ given that such t exists, and $T_j = \infty$ otherwise.

The definition for time-to-absorbption into a set $\mathcal{J} \subseteq \mathcal{A}$ is analogous, with $T_{\mathcal{J}} = \min\{t : X(t) \in \mathcal{J}\}$, given such t exists, and $T_{\mathcal{J}} = \infty$ otherwise.

Definition 46 (Cause-specific hazard function)**.**

Suppose that $||\mathcal{A}|| > 1$ (i.e. suppose that process $\{X(t)\}$ is associated with more than one absorbing state). We define the cause-specific hazard function associated with state (cause) $j \in \mathcal{A}$ given the process starts in state i as

$$\lambda_{ij}(t) = \lim_{h \to 0^+} \frac{P(t < T < t + h, X(T) = j \mid T > t, X(0) = i)}{h} = \frac{f_{ij}(t)}{1 - F_i(t)}, \tag{2.35}$$

where $f_{ij}(t)$ the probability density associated with T_j and $F_i(t)$ is the cumulative distribution function associated with T, each given that the process starts in state i.

The definition for the cause-specific hazard rate associated with a set $\mathcal{J} \subseteq \mathcal{A}$ is analogous,

$$\lambda_{i\mathcal{J}}(t) = \lim_{h \to 0^+} \frac{P(t < T < t + h, X(T) \in \mathcal{J} \mid T > t, X(0) = i)}{h} == \frac{f_{i\mathcal{J}}(t)}{1 - F_i(t)}, \tag{2.36}$$

where $f_{i\mathcal{J}}$ is the probability density function associated with $T_{\mathcal{J}}$ given that the process starts in state i.

The cause-specific hazard function is interpreted as the contribution to the hazard function associated with state j conditional on not having been absorbed into any $k \in \mathcal{A}$ before time t. In Chapter 3 we introduce a modified-cause-specific hazard rate, conditional only on not having been absorbed into some subset of $\mathcal{A}$, which is relevant to the evolution of gene duplicates.

By the analysis of the Markov chain, applying Proposition 4 we have

$$\begin{aligned} f_{ij}(t) &= \left[\underline{e}_i e^{\mathbf{Q}^* t} \mathbf{V}\right]_j \\ &= \underline{e}_i e^{\mathbf{Q}^* t} \underline{v}_j, \end{aligned} \tag{2.37}$$

where $\underline{v}_j$ is the j^{th} column of $\mathbf{V}$. Together with Theorem 9 (Theorem 2.4.1 in [72]), we have the following proposition.

Proposition 10.

For all $i \in \mathcal{S}^*$ and $j \in \mathcal{A}$

$$\lambda_{ij}(t) = \frac{\underline{e}_i e^{\mathbf{Q}^* t} \underline{v}_j}{\underline{e}_i e^{\mathbf{Q}^* t} \underline{\mathbf{1}}}. \tag{2.38}$$

The following proposition follows from the fact that events $\{X(t) = i\}$ and $\{X(t) = j\}$ are disjoint for $i \neq j$.

Proposition 11.

For all $i \in \mathcal{S}^*$, $\mathcal{J} \subseteq \mathcal{A}$, we have

$$\begin{aligned} f_{i\mathcal{J}}(t) &= \sum_{j \in \mathcal{J}} f_{ij}(t), \\ \lambda_{i\mathcal{J}}(t) &= \sum_{j \in \mathcal{J}} \lambda_{ij}(t), \\ f_i(t) &= \sum_{j \in \mathcal{A}} f_{ij}(t), \\ \lambda_i(t) &= \sum_{j \in \mathcal{A}} \lambda_{ij}(t). \end{aligned} \tag{2.39}$$

Definition 47 (Probability given initial distribution)**.**

We denote the probability of an event A given initial distribution $\underline{\alpha}_0$ by $P_{\underline{\alpha}_0}(A)$.

We denote the probability of an event A conditional on event B given initial distribution distribution $\underline{\alpha}_0$ by $P_{\underline{\alpha}_0}(A \mid B)$.

The following proposition follows immediately from the law of total probability

Proposition 12.

For any distribution $\underline{\alpha}_0$ and event A,

$$P_{\underline{\alpha}_0}(A) = \sum_{i \in \mathcal{S}} \underline{\alpha}_{0i} P(A | X(0) = i). \tag{2.40}$$

The next proposition establishes that $P_{\underline{\alpha}_0}(.)$ is analogous to $P(.)$ in terms of conditional probabilities.

Proposition 13.

For any distribution $\underline{\alpha}_0$ and events A, B

$$P_{\underline{\alpha}_0}(A \mid B) = \frac{P_{\underline{\alpha}_0}(A, B)}{P_{\underline{\alpha}_0}(B)}. \tag{2.41}$$

Proof.

Applying the law of total probability (and noting that B and $\{X(0) = i\}$ are not necessarily independent) we have

$$\begin{aligned} P_{\underline{\alpha}_0}(A \mid B) &= \sum_{i \in \mathcal{S}} P(A \mid B, X(0) = i) P(X(0) = i \mid B) \\ &= \sum_{i \in \mathcal{S}} \frac{P(A, B, X(0) = i) P(X(0) = i, B)}{P(B, X(0) = i) P(B)} \\ &= \frac{\sum_{i \in \mathcal{S}} P(A, B, X(0) = i)}{P(B)} \\ &= \frac{\sum_{i \in \mathcal{S}} P(A, B \mid X(0) = i) P(X(0) = i)}{P(B)} \end{aligned} \tag{2.42}$$

Applying the law of total probability to the denominator we have

$$\begin{aligned} P_{\underline{\alpha}_0}(A \mid B) &= \frac{\sum_{i \in \mathcal{S}} P(A, B \mid X(0) = i) P(X(0) = i)}{\sum_{i \in \mathcal{S}} P(B \mid X(0) = i) P(X(0) = i)} \\ &= \frac{\sum_{i \in \mathcal{S}} \alpha_{0i} P(A, B \mid X(0) = i)}{\sum_{i \in \mathcal{S}} \alpha_{0i} P(B \mid X(0) = i)} \end{aligned} \tag{2.43}$$

Finally, applying Proposition 12 we have

$$P_{\underline{\alpha}_0}(A \mid B) = \frac{P_{\underline{\alpha}_0}(A, B)}{P_{\underline{\alpha}_0}(B)}. \tag{2.44}$$

□

The following proposition combines Proposition 4 with the observations from [72; 87] (e.g. the first equation in the proof of Theorem 2.4.1 in [72] is enough).

Proposition 14 (Distribution over transient states)**.**

For an absorbing CTMC, the distribution over the transient states $\mathbf{P}^*(t) = [P_{ij}(t)]_{i,j \in \mathcal{S}^*}$ *is given by, for all* $t \geqslant 0$*,*

$$\mathbf{P}^*(t) = e^{\mathbf{Q}^* t}. \tag{2.45}$$

Definition 48 (Conditional transient distribution)**.**

We define $\underline{p}_i(t)$ *to be the distribution over the transient states conditional on the process not having been absorbed, and given start in state* $i \in \mathcal{S}^*$*. That is,* $\underline{p}_i(t) = [p_{ij}(t)]_{j \in \mathcal{S}^*}$ *where for all* $j \in \mathcal{S}^*$ *for each* $i \in \mathcal{S}^*$*,*

$$p_{ij}(t) = P(X(t) = j \mid X(t) \notin \mathcal{A}, X(0) = i). \tag{2.46}$$

When the initial distribution in the transient states is $\underline{\alpha}_0 = [\alpha_{0j}]_{j \in \mathcal{S}^*}$*, we denote the analogous distribution by* $\underline{p}_{\underline{\alpha}_0}(t)$*. That is* $\underline{p}_{\underline{\alpha}_0}(t) = [p_{\underline{\alpha}_{0j}}(t)]$ *where, for all* $j \in \mathcal{S}^*$*,*

$$p_{\underline{\alpha}_{0j}}(t) = P_{\underline{\alpha}_0}(X(t) = j \mid X(t) \notin \mathcal{A}). \tag{2.47}$$

The following proposition is stated without demonstration in [31]. We provide a proof here.

Proposition 15.

For all $t \geqslant 0$, and any initial distribution in the transient states $\underline{\alpha}_0$,

$$\underline{p}_{\underline{\alpha}_0}(t) = \frac{\underline{\alpha}_0 e^{\mathbf{Q}^* t}}{\underline{\alpha}_0 e^{\mathbf{Q}^* t}\underline{1}}. \tag{2.48}$$

Proof.

For each $j \in \mathcal{S}^*$, and all $t \geqslant 0$, we have

$$p_{\underline{\alpha}_{0j}}(t) = P_{\underline{\alpha}_0}(X(t) = j \mid X(t) \notin \mathcal{A}). \tag{2.49}$$

Applying Proposition 13 we have

$$\begin{aligned} p_{\underline{\alpha}_{0j}}(t) &= \frac{P_{\underline{\alpha}_0}(X(t) = j, X(t) \notin \mathcal{A})}{P_{\underline{\alpha}_0}(X(t) \notin \mathcal{A})} \\ &= \frac{P_{\underline{\alpha}_0}(X(t) = j)}{P_{\underline{\alpha}_0}(X(t) \notin \mathcal{A})}, \end{aligned} \tag{2.50}$$

since the events $\{(X(t) = j\}$ and $\{(X(t) = j) \cap (X(t) \notin \mathcal{A})\}$ are equivalent.

Now, applying Proposition 12, and noting that $P(X(t) \in \mathcal{S}^* \mid X(0) \notin \mathcal{S}^*) = 0$, we have

$$\begin{aligned} p_{\underline{\alpha}_{0j}}(t) &= \frac{\sum_{i \in \mathcal{S}^*} \alpha_{0i} P(X(t) = j \mid X(0) = i)}{\sum_{i \in \mathcal{S}^*} \alpha_{0i} P(X(t) \notin \mathcal{A} \mid X(0) = i)} \\ &= \frac{\sum_{i \in \mathcal{S}^*} \alpha_{0i} P_{ij}(t)}{\sum_{i \in \mathcal{S}^*} \alpha_{0i} P(T > t \mid X(0) = i)}. \end{aligned} \tag{2.51}$$

Now applying Proposition 14 and Theorem 9 we have

$$\begin{aligned} p_{\underline{\alpha}_{0j}}(t) &= \frac{\sum_{i \in \mathcal{S}^*} \alpha_{0i} \underline{e_i} e^{\mathbf{Q}^* t} \underline{e_j}}{\sum_{i \in \mathcal{S}^*} \alpha_{0j} \underline{e_i} e^{\mathbf{Q}^* \mathbf{t}} \underline{1}} \\ &= \frac{\underline{\alpha}_0 e^{\mathbf{Q}^* t} \underline{e_j}}{\underline{\alpha}_0 e^{\mathbf{Q}^* t} \underline{1}}, \end{aligned} \tag{2.52}$$

which in vector form is

$$p_{\underline{\alpha}_0}(t) = \frac{\underline{\alpha}_0 e^{\mathbf{Q}^* t}}{\underline{\alpha}_0 e^{\mathbf{Q}^* t}\underline{1}}. \tag{2.53}$$

□

Corollary 3.

For all $t \geqslant 0$,

$$\underline{p}_i(t) = \frac{\underline{e_i} e^{\mathbf{Q}^* t}}{\underline{e_i} e^{\mathbf{Q}^* t}\underline{1}}. \tag{2.54}$$

Definition 49 (Quasi-stationary distribution)**.**

A distribution $\underline{\alpha} = [\alpha_i]_{i \in \mathcal{S}^*}$ *is called a quasi-stationary distribution (QSD) if, for all* $t \geqslant 0$

$$\frac{\underline{p}_{\underline{\alpha}}(t)}{\underline{p}_{\underline{\alpha}}(t)\underline{1}} = \underline{\alpha}. \tag{2.55}$$

As shown by Darroch and Seneta [31], when the set of transient states is irreducible the quasi-stationary distribution is unique. We will be primarily concerned with the case that the QSD is unique. Van Doorn and Pollett [121] provide a comprehensive review of quasi-stationary distributions, including conditions for existence and uniqueness, and a guide for computing QSDs in MATLAB.

Definition 50 (Yaglom Limit)**.**

The Yaglom limit given initial distribution $\underline{\alpha}_0$ *is defined (given the limits exist) by* $\underline{y}_{\underline{\alpha}_0} = [y_j]_{j \in \mathcal{S}^*}$*, where for all* $j \in \mathcal{S}^*$*,*

$$y_j = \lim_{t \to \infty} \frac{P_{\underline{\alpha}_0}(X(t) = j)}{P_{\underline{\alpha}_0}(X(t) \notin \mathcal{A})} = \lim_{t \to \infty} P_{\underline{\alpha}_0}(X(t) = j \mid X(t) \notin \mathcal{A}). \tag{2.56}$$

The following observation (stated here as a proposition) follows immediately from Definitions 48 and 50.

Proposition 16.

For initial distribution $\underline{\alpha}_0$ *we have*

$$\underline{y}_{\underline{\alpha}_0} = \lim_{t \to \infty} \underline{p}_{\underline{\alpha}_0}(t). \tag{2.57}$$

Theorems 1 and 2 of Vere Jones [122] prove that the Yaglom limit is a quasi-stationary distribution. Thus, in the case of a unique QSD, the Yaglom limit is also unique, and we denote it by $\underline{y}$. We can interpret $y_i = \alpha_i$ as the long run probability that the process is in state i given that it has not been absorbed [31].

When the Yaglom limit depends on the initial distribution, it remains the case that every Yaglom limit is a QSD, and every QSD is a Yaglom limit for some initial distribution (namely, itself) [121]. We say that a QSD $\underline{\alpha}$ is associated with initial distribution $\underline{\alpha}_0$ when $\underline{\alpha} = \underline{y}_{\underline{\alpha}_0}$.

Definition 51 (Ratio of means distribution)**.**

We define the ratio of means distribution $\underline{\alpha}_1 = [\alpha_{1j}]_{j \in \mathcal{S}^*}$ *associated with initial distribution* $\underline{\alpha}_0$ *by, for all* $j \in \mathcal{S}^*$*,*

$$\alpha_{1j} = \frac{E_{\underline{\alpha}_0}(T_j^*)}{E_{\underline{\alpha}_0}(T)}, \tag{2.58}$$

or equivalently

$$\underline{\alpha}_1 = \frac{E_{\underline{\alpha}_0}(\underline{T}^*)}{E_{\underline{\alpha}_0}(T)}, \tag{2.59}$$

where T_j^* is a random variable tracking time spent in state $j \in \mathcal{S}^*$ before absorption, and $\underline{T}^* = [T_j^*]_{j\in\mathcal{S}^*}$.

The ratio of means distribution was introduced by Darroch and Seneta [30], but has received relatively little attention. Darroch and Seneta [30] noted that the ratio of means distribution is dependent on the initial distribution, while (in the context under discussion) the quasi-stationary is not. Artalejo and Lopez-Herrero [3] revived some interest in this distribution by noting that in certain contexts, dependence on the initial distribution is a desirable property. Often the initial distribution is known, and the assumption that the process is at equilibrium is not well justified, but analysis of long term behaviour is still pertinent.

Artalejo and Lopez-Herrero [3] discussed the examples of population biology, and in particular, epidemic models. In Chapter 5 we will see that this distribution also has applications in evolutionary biology.

Our final result for this section follows from Proposition 4, and Theorem 9, and was first shown by Darroch and Seneta [31]. Darroch and Seneta [30] also showed an equivalent result for the discrete-time case.

Theorem 11.

For any initial distribution $\underline{\alpha}_0$, the ratio of means distribution is given by

$$\begin{aligned}\underline{\alpha}_1 &= \frac{\int_0^\infty \underline{\alpha}_0 e^{\mathbf{Q}^* t}}{\int_0^\infty \underline{\alpha}_0 e^{\mathbf{Q}^* t}\underline{1}} \\ &= \frac{\underline{\alpha}_0(\mathbf{Q}^*)^{-1}}{\underline{\alpha}_0(\mathbf{Q}^*)^{-1}\underline{1}}.\end{aligned} \tag{2.60}$$

2.2 Likelihood and model selection

In this section we introduce the likelihood function, and the maximum likelihood estimator — a concept which sees a great deal of use in evolutionary biology (for a more detailed discussion see [98] or [17]). We then discuss some concepts from information theory which pertain to model selection in a maximum likelihood framework, particularly the Akaike- (AIC) and Bayesian- Information Criterion (BIC), which we make some use of in Chapter 5.

Generally in the model selection literature a continuous support is assumed. The models in this thesis have discrete support, but most of the results have obvious

discrete analogues, where density functions are replaced by probability mass functions, integrals by sums, etc. We have adapted the discussion of [17] for the setting with discrete support here.

Definition 52 (The likelihood function)**.**

Let $X = \{X_1, X_2, \ldots, X_n\}$ be a sequence of discrete random variables whose probability distribution $p_{\underline{\theta}}$ depends on the parameter $\underline{\theta} = [\theta_1, \theta_2, \ldots, \theta_m]$. Then we define a likelihood function by

$$L(\underline{\theta}; x) = p_{\underline{\theta}}(x) = P(X = x \mid \underline{\theta}). \tag{2.61}$$

We interpret likelihood function as the probability of seeing the data x given that the parameter $\underline{\theta}$ is the underlying parameter set for $p_{\underline{\theta}}$.

The following proposition follows immediately from the fact that $P(X = x, Y = y) = P(X = x)P(Y = y)$ for independent and identically distributed (iid) X, Y.

Proposition 17.

If $X_1, X_2, \ldots X_n$ are iid, then

$$L(\underline{\theta} \mid x) = \prod_{i=1}^{n} P(X_i = x_i \mid \underline{\theta}). \tag{2.62}$$

Definition 53 (Maximum likelihood estimator)**.**

The maximum likelihood estimator is a parameter $\hat{\underline{\theta}}$ such that,

$$L(\hat{\underline{\theta}}; x) = \sup_{\underline{\theta} \in \Theta} L(\underline{\theta}; x). \tag{2.63}$$

In practice it is often easier to compute using the log likelihood function, since taking products over many small terms is likely to result in numerical instability.

Proposition 18.

Since log *is a monotone function*

$$\log L(\hat{\underline{\theta}}; x) = \sup_{\underline{\theta} \in \Theta} \log L(\underline{\theta}; x). \tag{2.64}$$

Likelihood and the maximum likelihood estimator provides a sufficient framework for parameter selection within a model, however in comparing different models the likelihood alone is generally insufficient. Information theoretic approaches including the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) allow for likelihood-based model selection with appropriate penalties for over-parameterization [17].

Both the AIC and BIC can be conceptualized in terms of the Kullback–Leibler divergence, a measure of divergence from one probability distribution to another.

Definition 54 (Kullback–Leibler (KL) divergence)**.**

For probability mass functions p, $\hat{p}$, both with the underlying space $\mathcal{S}$, the KL divergence of $\hat{p}$ from p is defined as

$$D_{KL}(p||\hat{p}) = \sum_{x \in \mathcal{S}'} \log\left(\frac{p(x)}{\hat{p}(x)}\right) p(x),$$

where

$$\mathcal{S}' = \{x \in \mathcal{S} : p(x) \neq 0\} = \{x \in \mathcal{S} : \hat{p}(x) \neq 0\}.$$

If the above set equality does not hold, the KL divergence is undefined.

$D_{KL}(p||\hat{p})$ can be interpreted as the information lost when distribution $\hat{p}$ is used to approximate p. Note that D_{KL} is not symmetric, and does not obey the triangle inequality, and hence is not a metric.

Consider a set of models $\mathcal{M} = \{\mathcal{M}_j : j = 1, ..., k\}$ where

$$\mathcal{M}_j = \{p_j(x; \underline{\theta}) : \underline{\theta} \in \Theta_j\}, \tag{2.65}$$

with $p_j(x; \underline{\theta})$ a probability mass function, each with the same support $\mathcal{S}$, and the parameter space for model $\mathcal{M}_j$ is Θ_j.

Suppose that we wish to use a model in $\mathcal{M}$ to approximate some distribution f (also with support $\mathcal{S}$), which may or may not be in $\mathcal{M}_j$ for some j. Suppose also that we have some data $Y = \{Y_i : i = 1, ..., n\}$ drawn from the distribution f we wish to approximate. Let $\hat{\underline{\theta}}_j(Y) = \hat{\underline{\theta}}_j$ be the maximum likelihood parameter estimate for model j based on data Y.

We can measure the quality of an approximation to f from model $\mathcal{M}_j$ with parameter $\underline{\theta} \in \Theta_j$ using the KL divergence

$$\begin{aligned} D_{KL}(f||p_j(.; \underline{\theta})) &= \sum_{x \in \mathcal{S}} \log\left(\frac{f(x)}{p_j(x; \underline{\theta})}\right) f(x) \\ &= \sum_{x \in \mathcal{S}} \log(f(x)) f(x) - \sum_{x \in \mathcal{S}} \log(p_j(x; \underline{\theta})) f(x). \end{aligned} \tag{2.66}$$

Notice that only the second term of Equation (2.66) depends on the model $\mathcal{M}_j$ and parameter $\underline{\theta}$. Thus, we can choose a best (in the KL sense) model and parameter set to approximate distribution f by maximizing

$$K(j, \underline{\theta}) = \sum_{x \in \mathcal{S}} \log(p_j(x; \underline{\theta})) f(x). \tag{2.67}$$

Now consider some fixed j, and let $\underline{\theta}_0$ maximize $K(\underline{\theta}) = K(j, \underline{\theta})$, and we will omit j from subscripts and arguments from here. Suppose that we have some data $Y = \{Y_i : i = 1, ..., n\}$, which are independent and drawn from the distribution f we wish to approximate. Given sufficiently large n, we can apply the law of large numbers to write

$$f(x) \approx \frac{||\{Y_i = x\}||}{n},$$

from which it follows that

$$\begin{aligned} K(\underline{\theta}) &\approx \frac{1}{n} \sum_{x \in \mathcal{S}} \log(p(x; \underline{\theta}))||\{Y_i = x\}|| \\ &= \frac{1}{n} \sum_{i=1}^{n} \log(p(Y_i; \underline{\theta})) \\ &= \frac{1}{n} \log L(\underline{\theta}; Y). \end{aligned} \tag{2.68}$$

Clearly the right hand side of Equation (2.68) is maximised by the maximum likelihood estimator $\hat{\underline{\theta}}$ (since it's just the log likelihood divided by n), and $\hat{\underline{\theta}}$ converges in probability to $\underline{\theta}_0$. Hence $p(.; \hat{\underline{\theta}})$ estimates the best-fitting distribution (in terms of KL divergence) in the model to the true distribution $f(.)$, or $\hat{\underline{\theta}}$ estimates $\underline{\theta}_0$.

However, we have used Y to estimate the true distribution f in the above; if we were also to use Y to attain a maximum likelihood estimate our estimated distribution would clearly be biased. Akaike [1] noted that this bias could be eliminated by instead minimizing the expected KL divergence.

Making some approximations, he showed that if the model was correct, in the sense that for some $\underline{\theta}$, $p(x; \underline{\theta}) = f(x)$ for all x, (or almost everywhere in the case of a continuous support), this leads to minimising

$$\hat{K}(\underline{\theta}) = \frac{1}{n}(\log L(\underline{\theta}; Y) - k), \tag{2.69}$$

where $k = \dim(\underline{\theta})$. Multiplying Equation (2.69) by $2n$ (for historical reasons — this has no effect on the procedure) we get Akaike's information criterion

$$\text{AIC} = 2k - 2 \log L(\underline{\theta}; Y). \tag{2.70}$$

Thus minimising AIC allows us to pick (approximately) the KL-best model from among a set of models, and within each model the parameter set which minimises the AIC will be the maximum likelihood estimator.

The Bayesian Information Criterion (BIC) introduced by Schwarz [103] is similar to the AIC,

$$\text{BIC} = \log(n)k - 2 \log L(\underline{\theta}; Y). \tag{2.71}$$

As the name suggests, the underlying philosophy of the BIC is a Bayesian approach to model selection. Kuha [69] provides a discussion of the relative merits of the AIC and BIC. For our purposes in Chapter 5, we use the BIC over the AIC because of its steeper penalty for overparameterization.

Chapter 3

Duplicate Pairs Evolving Under Subfunctionalization

This chapter is the topic of my recent paper [109] with Assoc. Prof. David A. Liberles, Assoc. Prof. Barbara R. Holland and Dr. Małgorzata M. O'Reilly. Compared to the paper (which has a more biological focus), we focus particularly on the mathematical results. Several results which do not appear in the paper (or its appendices) are included in this chapter.

Here we consider in detail the evolution of a duplicate pair immediately after some duplication event. Recall from Section 1.2 of the Introduction that subfunctionalization is a process by which the functions of an ancestral gene are distributed among two (or more, if multiple duplication events occur) duplicate copies of the original gene. We assume throughout that the duplication event led to two perfect copies of the duplicated gene, and that no further duplication occurs; this is equivalent to considering the evolution of a *gene family* of fixed size $n = 2$.

To explain the subfunctionalization process, we can think of a gene as being divided into regions of two types. The first type is the coding region which we can think of as a single unit which must be maintained in order for the gene as a whole to function. The other type are the regulatory regions, of which there are at least two for any gene which undergoes subfunctionalization. We think of the regulatory regions as corresponding to some function of the gene. When a gene is duplicated, it creates some redundancy in the regulatory regions — we say that a regulatory region which is maintained in both copies of the gene is redundant, and immediately after duplication this is the case for all regulatory regions.

We assume that the maintenance of at least one copy of each regulatory region is

essential (an organism without a functioning copy would die, and hence is not represented in the population). Redundant regulatory regions in each copy fail at a Poisson rate u_r. Furthermore, if the coding region is susceptible to failure (if and only if all of its associated functioning regulatory regions are redundant), then it does so at Poisson rate u_c.

Figure 3.1 illustrates the full set of possible realisations of the process, up to the order of the regulatory regions, for a duplicate pair with four regulatory regions.

The assumption that the two copies are perfect can be trivially relaxed by considering initial states besides $X(0) = 0$ in the Markov chain introduced below, but we do not treat this case explicitly. We are also interested in relaxing the assumption that no further duplication occurs. In Section 4.2 we discuss the intuitions behind modelling the case in which there are more than two duplicates, and where further duplication is allowed. We also formally present a model for gene families of size $n > 2$, but this model still treats the number of duplicates as fixed.

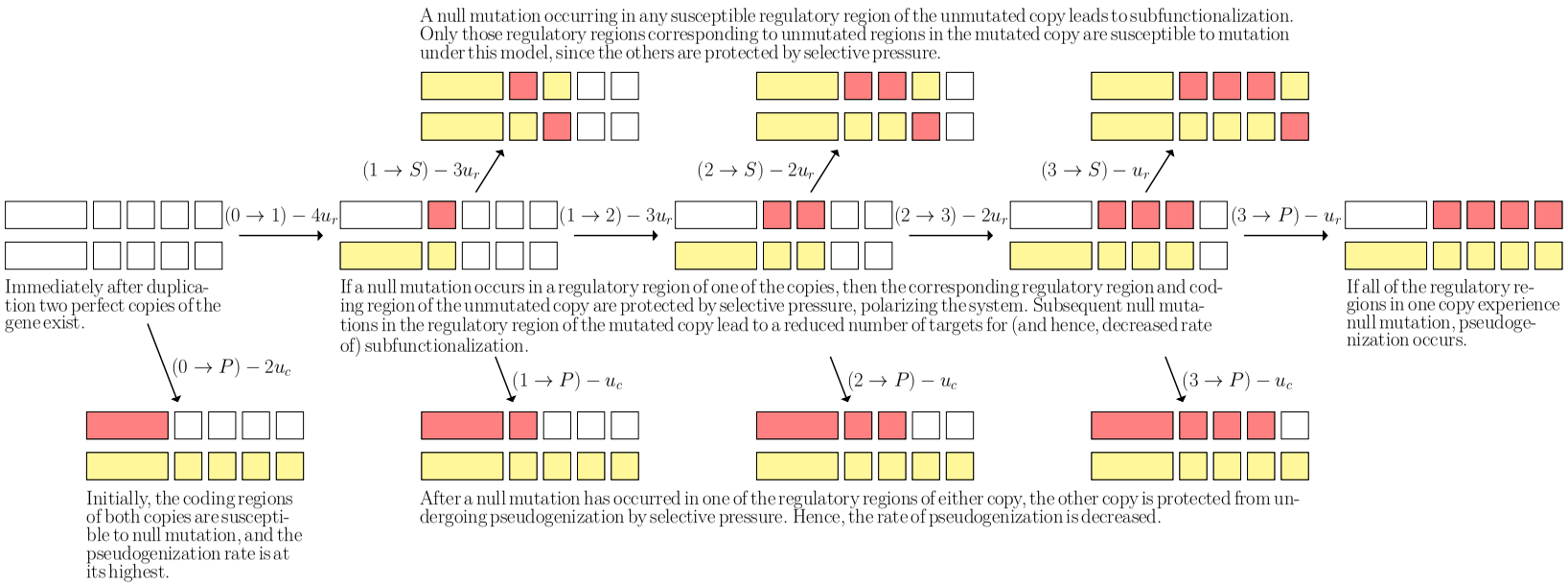

Figure 3.1: The (biological) transition diagram for $z = 4$. Regions hit by null mutation are coloured red, and regions which are protected by selective pressure are coloured yellow.

3.1 A model for the evolution of a pair of gene duplicates

Consider the evolution of two copies of a duplicate gene, evolving according to the duplication-degeneration-complementation (DDC) process [42]. Under this process, the subfunctions of the original gene are divided up among the two copies such that each function is always preserved in at least one gene. Initially, both copies have functional copies of each regulatory region, and are thus associated with all of original gene's subfunctions. Over time, the copies can experience mutations in their regulatory regions leading to the loss of a particular subfunction. Usually such a mutation would be deleterious, and strongly selected against, but the redundancy created by the presence of two copies allows for such mutations to occur without negative selection. A mutation which is not (or at least not significantly) selected against is called neutral.

If a gene is uniquely associated with any subfunction, there is no such redundancy, and it is protected against pseudogenization by selective pressure, such that the duplicate pair can ultimately reach one of two fates — either both copies become uniquely associated with (separate) subfunctions, and both are preserved (subfunctionalization), or one copy ultimately loses all functionality (pseudogenization). Immediately after duplication, we assume that both copies each have z mutable regulatory regions, with each regulatory region associated with a single (see Remark 3) unique subfunction. We assume that null mutations become fixed in the regulatory regions at some Poisson rate u_r which is identical for each regulatory region in each of the two copies. We further assume that null mutations fix in the coding regions of each copy at a rate u_c (we think of the coding region of each copy as being a single unit).

Remark 3.

The assumption that each regulatory region is associated with a single subfunction can be relaxed by modeling a single regulatory region associated to multiple subfunctions as two or more separate regulatory regions associated to a single subfunction. We do not treat this explicitly, but it is worth keeping in mind.

For a fixed number z of the regulatory regions in the duplicate pair of genes, consider a continuous-time Markov chain $\{X(t), t \geqslant 0\}$, with state space

$$\mathcal{S} = \{0, 1, ..., z-1\} \cup \{S, P\}, \tag{3.1}$$

where state $i \in \{0, 1, ..., z-1\}$ represents the number of fixed null mutations to have occurred in the case that neither subfunctionalization nor pseudogenization have hap-

pened yet, and the states S and P are introduced to represent subfunctionalization and pseudogenization respectively. S and P are both absorbing states — that is, once subfunctionalization or pseudogenization occurs, the process stops and remains in state S or P, which represent the preservation of both copies, or one copy respectively. Under the subfunctionalization process, a duplicate pair is preserved if it undergoes subfunctionalization, otherwise one gene is lost (pseudogenization) and the remaining gene is preserved.

Note that we have simplified the problem by modeling the number of null mutations to have occurred in the system as whole, rather than tracking mutations in each gene separately. This simplification does not result in any loss of information, because as soon as a null mutation has occurred in both genes, either subfunctionalization or pseudogenization must have occurred, and as such we need only count the total number of mutations until one of these two possible outcomes is realized.

We define the generator for our Markov chain to be matrix $\mathbf{Q} = [q_{ij}]$ where the non-zero off-diagonals are given by

$$q_{ij} = \begin{cases} 2u_c & \text{if } i = 0, j = P \\ 2zu_r & \text{if } i = 0, j = 1 \\ u_c & \text{if } 1 \leqslant i \leqslant z-2, j = P \\ (z-i)u_r & \text{if } 1 \leqslant i \leqslant z-2, j = i+1 \text{ or } j = S \\ u_r + u_c & \text{if } i = z-1, j = P \\ u_r & \text{if } i = z-1, j = S. \end{cases} \tag{3.2}$$

Below, we show that the rates q_{ij} in (3.2) are indeed the relevant transition rates by considering the evolution immediately after duplication.

Transitions from $0 \to P$

Clearly, the process starts in state 0, since no null mutations have fixed at the instant of duplication. Null mutations fix in the coding region for each gene at a rate u_c, and this leads to pseudogenization. Therefore transitions from $0 \to P$ occurs at rate $2u_c$.

Transitions from $0 \to 1$

Null mutations fix in each of the $2z$ regulatory regions at a rate u_r, and hence transition $0 \to 1$ occurs at a rate $2zu_r$.

After the first mutation, either a null mutation fixed in one of the coding regions, and the process has been absorbed into state P, or a null mutation has fixed in one of the regulatory regions of one of the genes, and the process is now in state 1.

As described in [81], null mutation in the regulatory region results in the loss of some

particular function for that gene, and the total loss of a function is selected against. Hence the duplicate pair must retain at least one unmutated copy of each regulatory region between them — this is the fundamental concept of subfunctionalization. It follows then that the remaining unaffected gene must be preserved, and so too must its copy of the regulatory region which has mutated in the other duplicate.

Transitions from $1 \to P$

Since the unaffected gene now has a unique function which is protected by selective pressure, this gene is no longer susceptible to pseudogenization under the subfunctionalization process. As such, only one copy may now undergo null mutation in the coding region, which it does at a rate u_c. Hence the rate of transitions from $1 \to P$ is u_c.

Transitions from $1 \to S$

Also, since one regulatory region in the unaffected gene is protected by selective pressure, and one region has already undergone null mutation for the other gene, each gene has $z-1$ regulatory regions which are now susceptible to null mutation. If such a mutation occurs in the previously unaffected copy, then both copies will have a unique function, and both will be protected by selective pressure. This is subfunctionalization, and hence the process transitions from $1 \to S$ at a rate $(z-1)u_r$.

Transitions from $1 \to 2$

On the other hand, if a null mutation fixes in one of the $z-1$ susceptible regulatory regions of the same copy in which the previous mutation fixed then the process transitions to state 2 — as two mutations have now fixed, but the process has not yet been absorbed. Hence the process transitions from $1 \to 2$ at a rate $(z-1)u_r$.

Transitions from $i \in \{1, 2, .., z-2\} \to j$

We note that for the process to reach state $i \in 1, 2, ..., z-2$ all mutations must have occurred in the regulatory regions of the same copy, since subfunctionalization (and hence absorption to S) occurs as soon as both copies have a unique function. Therefore, a similar argument is used to show that for all $i \in \{1, 2, .., z-2\}$ transitions

- from $i \to P$ occur at rate u_c;
- from $i \to S$ occur at rate $(z-i)u_r$; and
- from $i \to i+1$ occur at rate $(z-i)u_r$.

Transitions from $z-1 \to S$

When the process is in state $z-1$ there is only one regulatory region for each copy susceptible to null mutation, which occurs at a rate u_r for each copy. If such a mutation

occurs in the so-far unaffected gene then the process transitions to state S, hence the rate of transition from $z-1 \to S$ is u_r.

Transitions from $z-1 \to P$

There are two distinct ways in which the process can transition from $z-1 \to P$. The first is similar to the previous cases, with a null mutation occurring in the coding region of the copy in which all of the previous mutations have fixed, which occurs at rate u_c. The other way is for a null mutation to fix in the last remaining regulatory region of this same gene, which occurs at a rate u_r. Hence the rate of transition from $z-1 \to P$ is $u_r + u_c$.

The full set of possible transitions for the case where $z = 4$ is illustrated in Figure 3.1.

Remark 4.

For fixed z, considered as a function of u_r and u_c we have $k\mathbf{Q}(u_r, u_c) = \mathbf{Q}(ku_r, ku_c)$. Thus in the absence of a relative clock (such as synonymous site mutations, which we use in our data analysis) the parameters u_c and u_r are only meaningful relative to each other. As such, we can replace parameters u_c and u_r with their ratio $\gamma = u_r/u_c$, and work with time units $1/u_c$. We apply this technique implicitly at various points in the discussion which follows.

Biologically likely parameter sets

In the gene duplication literature, it is generally supposed that the rate of null mutation in the coding region most often significantly exceeds that in the associated regulatory regions [42; 83; 54]. It is also generally thought that most genes have just a few regulatory regions, with more than 10 being regarded as quite a large number. These are not well established observations, but rather represent a combination of expert intuition, and a relatively small amount of empirical observation (e.g. [65; 2]).

As such, we will pay special interest to parameter sets with $u_r < u_c$, though we do not neglect the case with $u_r \geqslant u_c$ entirely, and we focus on $z \in \{2, ...16\}$. In our data analysis section (3.8) we establish some evidence in favour of this, finding that z is most likely in the range from 3 to 5, and that u_r is most likely 5 to 10 times smaller than u_c (for four species).

3.2 Probabilities corresponding to the i^{th} mutational events

Here we derive probabilities P_i^{*z} and S_i^{*z} corresponding to the i^{th} mutational events at which absorption into either pseudogenization or subfunctionalization occurs, respectively. Some of these results have previously been discussed by Force et al. [42], here we will derive them from a Markov chain analysis.

For a fixed number z of regulatory regions in the duplicate pair of genes, consider a discrete-time Markov chain $\{X_n, n = 1, 2, \ldots\}$, whose states are observed at the n^{th} mutational events, with state space

$$\mathcal{S} = \{0, 1, ..., z-1\} \cup \{S, P\}, \tag{3.3}$$

where state $i \in \{0, 1, ..., z-1\}$ represents the number of fixed null mutations to have occurred in the case that neither subfunctionalization nor pseudogenization have happened yet, and the states S and P are introduced to represent subfunctionalization and pseudogenization respectively. S and P are both absorbing states, since under the subfunctionalization model a duplicate pair is preserved if it undergoes subfunctionalization, otherwise one gene is lost (pseudogenization) and the remaining gene is preserved. Note that this is the embedded chain [100] of the continuous-time Markov chain with generator given in Equation (3.2).

We assume that the initial state at time zero is $X_0 = 0$ (that is, we assume that the process starts with two perfect copies of the duplicated gene). For $i = 1, \ldots, z$, define

- $P_i^z = P(X_i = P \mid X_k \notin \{S, P\}, k = 1, \ldots, i-1) = P(X_i = P \mid X_{i-1} = z - i + 1)$, interpreted as the probability of pseudogenizing at the i^{th} mutational event, given that neither pseudogenization nor subfunctionalization has occurred yet;
- $S_i^z = P(X_i = S \mid X_k \notin \{S, P\}, k = 1, \ldots, i-1) = P(X_i = S \mid X_{i-1} = z - i + 1)$, interpreted as the probability of subfunctionalizing at the i^{th} mutational event, given that neither pseudogenization nor subfunctionalization has occurred yet;
- $P_i^{*z} = P(X_i = P, X_k \neq P, k = 1, \ldots, i-1)$, interpreted as the probability of pseudogenizing at the time of the i^{th} mutational event; and
- $S_i^{*z} = P(X_i = S, X_k \neq S, k = 1, \ldots, i-1)$, interpreted as the probability of subfunctionalizing at the time of the i^{th} mutational event.

By the analysis of the Markov chain, with P_i^z evaluated as in [53], we have $S_1^z = 0$,

$$P_1^z = \frac{u_c}{u_c + zu_r}, \tag{3.4}$$

$$P_z^z = \frac{u_c + u_r}{u_c + 2u_r}, \tag{3.5}$$

and, for $2 \leqslant i \leqslant z-1$,

$$P_i^z = \frac{u_c}{u_c + 2(z-i+1)u_r}; \tag{3.6}$$

for $2 \leqslant i \leqslant z$,

$$S_i^z = \frac{(z-i+1)u_r}{u_c + 2(z-i+1)u_r}; \tag{3.7}$$

for $1 \leqslant i \leqslant z$,

$$P_i^{*z} = \prod_{k=1}^{i-1}(1 - P_k^z - S_k^z) \cdot P_i^z, \tag{3.8}$$

and

$$P_{i+1}^{*z} = (1 - P_i^z - S_i^z) \cdot \frac{P_{i+1}^z}{P_i^z} \cdot P_i^{*z}; \tag{3.9}$$

for $3 \leqslant i \leqslant z$,

$$S_i^{*z} = \prod_{k=1}^{i-1}(1 - P_k^z - S_k^z) \cdot S_i^z, \tag{3.10}$$

and

$$S_{i+1}^{*z} = (1 - P_i^z - S_i^z) \cdot \frac{S_{i+1}^z}{S_i^z} \cdot S_i^{*z}; \tag{3.11}$$

and also $S_1^{*z} = 0$, $S_2^{*z} = (1 - P_1^z - S_1^z) \cdot S_2^z$.

Clearly, since absorption into $\{S, P\}$ must occur by the time of the z^{th} mutational event with probability 1, we also have

$$\sum_{i=1}^{z}(P_i^{*z} + S_i^{*z}) = 1, \tag{3.12}$$

and note that the quantity F_i^{*z} defined as

$$F_i^{*z} = \sum_{k=1}^{i}(P_k^{*z} + S_k^{*z}), \tag{3.13}$$

is the probability of having been absorbed into $\{S, P\}$ at or before the time of the i^{th} mutational event.

Further, denote by T_i the random variable recording the time of the i^{th} mutational event, let $\Delta T_i = T_i - T_{i-1}$, and note that, as in [53],

$$E(\Delta T_i) = \begin{cases} \frac{1}{u_c + 2(z-i)u_r} & \text{if } 1 \leqslant i \leqslant z \\ \frac{1}{2(u_c + zu_r)} & \text{if } i = 1 \\ \frac{1}{u_c + 2u_r} & \text{if } i = z. \end{cases} \tag{3.14}$$

3.3 Hazard function and related measures

3.3.1 Hazard function

Let

$$T_{\{S,P\}} = \inf\{t > 0 : X(t) \in \{S, P\}\}, \tag{3.15}$$

be the time at which the absorption into $\{S, P\}$ occurs. Following the definition in (2.31), the hazard rate at time t for absorption into $\{S, P\}$, given that the process starts in state $i \in \{0, 1, \ldots, z-1\}$, is given by

$$\lambda_i(t) = \lim_{h \to 0^+} \frac{P(t < T_{\{S,P\}} < t + h \mid T_{\{S,P\}} > t, X(0) = i)}{h} = \frac{f_i(t)}{1 - F_i(t)}, \tag{3.16}$$

where $f_i(t)$ is the probability density of absorption occurring at time t given start in state i, and

$$F_i(t) = \int_{u=0}^{t} f_i(u)du, \tag{3.17}$$

is the corresponding cumulative distribution function. The rate (3.16) measures the instantaneous rate of absorption into any absorbing state, given that the process has not yet been absorbed. That is the *hazard rate corresponding to absorption into* $\{S, P\}$.

From Theorem 10 in Section 2.1.3 we have,

$$\lambda_i(t) = \frac{-\underline{e_i} e^{\mathbf{Q}^* t} \mathbf{Q}^* \underline{\mathbf{1}}}{\underline{e_i} e^{\mathbf{Q}^* t} \underline{\mathbf{1}}}. \tag{2.33}$$

3.3.2 Cause-specific hazard rates

When an absorption into $\{S, P\}$ occurs, the process transitions to either S or P. Recalling Equation (2.35) the cause-specific hazard rate is given by,

$$\begin{aligned} \lambda_{ij}(t) &= \lim_{h \to 0^+} \frac{P(t < T_{\{S,P\}} < t + h, X(T_{\{S,P\}}) = j \mid T_{\{S,P\}} > t, X(0) = i)}{h} \\ &= \frac{f_{ij}(t)}{1 - F_i(t)}, \end{aligned} \tag{2.35}$$

where $f_{ij}(t)$ is the probability density function associated with random variable T_j, i.e. the probability density of absorption occurring at time t *and* the absorption occurring into the specific state j, given start in state i. Note that $f_{ij}(t)$ includes point mass at $t = \infty$ corresponding to the possibility that absorption into $k \neq j$ occurs. $F_i(t)$ is the cumulative distribution function associated with time to absorption into any state, given start in state i.

From Proposition 10 in Section 2.1.3 we have

$$\lambda_{ij}(t) = \frac{\underline{e_i} e^{\mathbf{Q}^* t} \underline{v_j}}{\underline{e_i} e^{\mathbf{Q}^* t} \underline{\mathbf{1}}}. \tag{2.38}$$

By the application of Lemma 1, which we prove in a more general form below, we have,

$$\lim_{t\to\infty} \lambda_{ij}(t) = \begin{cases} u_r + u_c & \text{for } j = P \\ u_r & \text{for } j = S. \end{cases} \tag{3.18}$$

So the cause-specific hazard rates $\lambda_{iS}(t)$ and $\lambda_{iP}(t)$ converge to u_r and $u_r + u_c$ respectively, as $t \to \infty$, and it follows from Corollary 4 (below) that

$$\lim_{t\to\infty} \lambda_i(t) = 2u_r + u_c. \tag{3.19}$$

Lemma 1.

Let $X(t)$ be an absorbing CTMC with some finite state space $\mathcal{S} = \{1, ..., m\} \cup \mathcal{A}$, with $\mathcal{A}$ being the set of absorbing states, and generator $\mathbf{Q} = [q_{ij}]_{i,j\in S}$ such that $q_{ij} = 0$ for all $j < i, i, j \in \{1, ..., m\}$, and state m being accessible from any $i \in \{1, ..., m-1\}$. Then, for any $i \in \{1, ..., m\}, j \in \mathcal{A}$,

$$\lim_{t\to\infty} \lambda_{ij}(t) = q_{mj}. \tag{3.20}$$

Proof.

Let $T_i' = \inf\{t > 0 : X(t) = i\}$ be the first time state i is visited, and denote the events $A_t^h = \{t < T_j < t + h\}$, $B_t = \{T_j > t\}$, $C_t = \{T_m' \leqslant t\}$, $\overline{C}_t = \{T_m' > t\}$.

By the law of total probability and the memoryless property of the Markov chain, and since

$$\lim_{t\to\infty} P(C_t \mid B_t, X(0) = i) = 1 \quad \text{and} \quad \lim_{t\to\infty} P(\overline{C}_t \mid B_t, X(0) = i) = 0, \tag{3.21}$$

it follows that the limits of $\lambda_{ij}(t)$ as $t \to \infty$ are

$$\begin{aligned}
\lim_{t\to\infty} \lambda_{ij}(t) &= \lim_{t\to\infty} \lim_{h\to 0^+} \frac{P(A_t^h, X(T) = j \mid B_t, X(0) = i)}{h} \\
&= \lim_{t\to\infty} \lim_{h\to 0^+} \frac{P(A_t^h, X(T) = j \mid B_t, X(0) = i, C_t) P(C_t \mid B_t, X(0) = i)}{h} \\
&\quad + \lim_{t\to\infty} \lim_{h\to 0^+} \frac{P(A_t^h, X(T) = j \mid B_t, X(0) = i, \overline{C}_t) P(\overline{C}_t \mid B_t, X(0) = i)}{h} \\
&= \lim_{t\to\infty} \left[P(C_t \mid B_t, X(0) = i) \lim_{h\to 0^+} \frac{P(A_t^h, X(T) = j \mid B_t, X(0) = i, C_t)}{h} \right] \\
&\quad + \lim_{t\to\infty} \left[P(\overline{C}_t \mid B_t, X(0) = i) \lim_{h\to 0^+} \frac{P(A_t^h, X(T) = j \mid B_t, X(0) = i, \overline{C}_t)}{h} \right] \\
&= \lim_{t\to\infty} \lim_{h\to 0^+} \frac{P(A_t^h, X(T) = j \mid B_t, X(0) = i, C_t)}{h} \\
&= \lim_{t\to\infty} \lim_{h\to 0^+} \frac{P(t < T < t + h, X(T) = j \mid (T > t, X(0) = i, T'_m \leqslant t)}{h} \\
&= \lim_{t\to\infty} \lim_{h\to 0^+} \frac{P(t < T < t + h, X(T) = j \mid X(t) = m, X(0) = i)}{h} \\
&= \lim_{t\to\infty} \lim_{h\to 0^+} \frac{P(t < T < t + h, X(T) = j \mid X(t) = m)}{h} \\
&= \lim_{t\to\infty} \lim_{h\to 0^+} \frac{P(X(t + h) = j \mid X(t) = m)}{h} \\
&= \lim_{h\to 0^+} \frac{P(X(h) = j \mid X(0) = m)}{h} \\
&= q_{m,j}.
\end{aligned}$$

□

Crossman [28] showed a similar result for hazard rates where the set of transient states is irreducible, and the process is a birth-death process, with transitions to the absorbing state occurring from precisely one transient state. We have an analogue of this result for the hazard rate of a process for which transitions between transient sets follow a pure-birth process, and transitions to the absorbing state can occur from arbitrarily many transient states. This follows as a corollary of Lemma 1 (applying Proposition 11 from Section 2.1.3).

Corollary 4.

For all $i \in \mathcal{S}\backslash\mathcal{A}$, and $\mathcal{J} \subseteq \mathcal{A}$, we have

$$\lim_{t\to\infty} \lambda_{i\mathcal{J}}(t) = \sum_{j\in\mathcal{J}} q_{mj}, \tag{3.22}$$

and

$$\lim_{t\to\infty} \lambda_i(t) = \sum_{j\in\mathcal{A}} q_{mj}. \tag{3.23}$$

3.3.3 Pseudogenization rate and survival function

In the context of subfunctionalization, the hazard rate and cause-specific hazard rates fail to capture an important feature of the model. The subfunctionalization state is better thought of as providing immunity to subsequent pseudogenization than as an alternate fail-state, but detecting whether subfunctionalization has occurred for particular duplicate pairs is not always possible in practice. As such, we are interested in the rate of pseudogenization given only that pseudogenization has not occurred, allowing for the possibility that subfunctionalization may have occurred — to that end, we introduce a new take on the hazard rate below, and prove a few general results which will be useful throughout our analysis of the subfunctionalization process.

Definition 55 (Modified-cause-specific hazard function)**.**

Let $\{X(t)\}$ be a CTMC with at least two absorbing states, state space $\mathcal{S}$, and set of absorbing states $\mathcal{A}$.

We define the modified-cause-specific hazard function associated with state $j \in \mathcal{A}$ given the process starts in state i as

$$\lambda_i^j(t) = \lim_{h \to 0^+} \frac{P(t < T_j < t + h \mid T_j > t, X(0) = i)}{h} = \frac{f_{ij}(t)}{1 - F_{ij}(t)}, \tag{3.24}$$

where $f_{ij}(t)$ and $F_{ij}(t)$ are (respectively) the density and cumulative distribution functions associated with random variable T_j.

Notice that the modified-cause-specific hazard function is conditional only on the process not having been absorbed into state j — absorption into $k \in \mathcal{A}, k \neq j$ is accounted for, and this distinguishes $\lambda_i^j(t)$ from the cause-specific hazard $\lambda_{ij}(t)$.

Remark 5.

Unless otherwise stated, it can be assumed for the rest of this chapter that we are discussing a CTMC as given in Definition 55 above.

Proposition 19.

For all $i \in \mathcal{S}, j \in \mathcal{A}, t \geqslant 0$,

$$\lambda_i^j(t) = \frac{\underline{\mathbf{e_i}} e^{\mathbf{Q}^* t} \underline{v}_j}{1 - \underline{\mathbf{e_i}} \left(e^{\mathbf{Q}^* t} - \mathbf{I}\right) \left(\mathbf{Q}^*\right)^{-1} \underline{v}_j}. \tag{3.25}$$

Proof.

By assumption, there exists $j, k \in \mathcal{A}$ with $k \neq j$. Let $\mathcal{A}^* = \mathcal{A} \backslash \{j\}$, and T_j be a random variable tracking time to absorption into state j. T_j has associated density function

$$f_{ij}(t) = \underline{\mathbf{e_i}} e^{\mathbf{Q}^* t} \underline{v}_j, \tag{3.26}$$

where $\underline{v}_j$ is the vector of transition rates into state j (i.e. it's the j^{th} column of matrix $\mathbf{V}$).

The expression for $F_{ij}(t)$ follows as

$$\begin{aligned} F_{ij}(t) &= \int_0^t f_{ij}(u)du \\ &= \int_0^t \underline{\mathbf{e_i}} e^{\mathbf{Q}^* u} \underline{v}_j du \\ &= \left[\underline{\mathbf{e_i}} e^{\mathbf{Q}^* u} (\mathbf{Q}^*)^{-1} \underline{v}_j \right]_0^t \\ &= \underline{\mathbf{e_i}} \left(e^{\mathbf{Q}^* t} - \mathbf{I}) \right) (\mathbf{Q}^*)^{-1} \underline{v}_j, \end{aligned} \tag{3.27}$$

hence,

$$\lambda_i^j(t) = \frac{\underline{\mathbf{e_i}} e^{\mathbf{Q}^* t} \underline{v}_j}{1 - \underline{\mathbf{e_i}} \left(e^{\mathbf{Q}^* t} - \mathbf{I}) \right) (\mathbf{Q}^*)^{-1} \underline{v}_j}. \tag{3.28}$$

□

The result below follows immediately from the fact that

$$\lim_{t \to \infty} e^{Q^* t} = 0.$$

Corollary 5.

For all $i \in \mathcal{S}, j \in \mathcal{A}$,

$$\lim_{t \to \infty} \lambda_i^j(t) = 0. \tag{3.29}$$

Lemma 2.

For all $i \in \mathcal{S}, j \in \mathcal{A}, n \in \mathbb{N}$ the n^{th} derivative of $\lambda_i^j(t)$ (which is clearly infinitely differentiable) approaches 0 as $t \to \infty$ i.e.

$$\lim_{t \to \infty} (\lambda_i^j(t))^{(n)} = 0. \tag{3.30}$$

Proof.

By the general Leibniz rule [90, p. 318] for any $n \in \mathbb{N}$ we have, by Equation (3.24),

$$(\lambda_i^j(t))^{(n)} = (f(1-F)^{-1})^{(n)}(t) = \sum_{k=0}^{n} \binom{n}{k} f^{(n-k)}(t)((1-F)^{-1})^{(k)}(t). \tag{3.31}$$

From basic properties of the matrix exponential we know,

$$(e^{\mathbf{Q}^* t})^{(i)} = e^{\mathbf{Q}^* t} (\mathbf{Q}^*)^i,$$

so,

$$f^{(n-k)}(t) = \underline{\mathbf{e_i}} e^{\mathbf{Q}^* t} (\mathbf{Q}^*)^{n-k} \underline{v}_j.$$

Thus

$$\lim_{t\to\infty} f^{(n-k)}(t) = 0, \text{ for all } (n-k) \in \mathbb{N}. \tag{3.32}$$

It follows that the first term (with $k = 0$) on the right hand side of Equation (3.31) is 0 in the limit as $t \to \infty$.

For the other terms, we need to consider $((1-F)^{-1})^{(k)}(t)$, which is more complicated, itself requiring application of Leibniz's rule to evaluate. However, Theorem 1 of [76] proves that

$$((1-F)^{-1})^{(k)}(t) = \sum_{i=1}^{k} (-1)^k \binom{k+1}{i+1} \frac{(1-F)^{(i)}(t)}{(1-F)^{i+1}(t)}.$$

Then since $(1-F(t))^{(i)} = f^{(i-1)}(t)$, for all $i \in \mathbb{N}^+$ and $(1-F(t))^{i+1} \geqslant \delta$ for some $\delta > 0$ as $t \to \infty$ for all $i \in \mathbb{N}$, we have

$$\lim_{t\to\infty} ((1-F)^{-1})^{(n)}(t) = 0, \text{ for all } n \in \mathbb{N}^+. \tag{3.33}$$

Taking the limit as $t \to \infty$ in Equation (3.31), and noting that the first term is 0 in the limit as $t \to \infty$, we have

$$\begin{aligned}
\lim_{t\to\infty} (\lambda_i^j(t))^{(n)} &= \lim_{t\to\infty} \sum_{k=0}^{n} \binom{n}{k} f^{(n-k)}(t)((1-F)^{-1})^{(k)}(t) \\
&= \sum_{k=0}^{n} \binom{n}{k} \lim_{t\to\infty} f^{(n-k)}(t)((1-F)^{-1})^{(k)}(t) \\
&= \sum_{k=1}^{n} \binom{n}{k} \lim_{t\to\infty} f^{(n-k)}(t)((1-F)^{-1})^{(k)}(t) \\
&= \sum_{k=1}^{n} \binom{n}{k} \lim_{t\to\infty} f^{(n-k)}(t) \lim_{t\to\infty} ((1-F)^{-1})^{(k)}(t).
\end{aligned}$$

Now substituting Equations (3.32) and (3.33) we get

$$\begin{aligned}
\lim_{t\to\infty} (\lambda_i^j(t))^{(n)} &= \sum_{k=1}^{n} \binom{n}{k} 0 \times 0 \\
&= 0.
\end{aligned}$$

□

In modelling the subfunctionalization process we are particularly interested in the pseudogenization modified-cause-specific hazard rate given start in state $i = 0$ (i.e. starting with two perfect copies) $\lambda_0^P(t)$, which we will denote by $h(t)$ and refer to as

the pseudogenization rate for convenience. In this case, Equations (3.24) and (3.25) together become

$$\begin{aligned} h(t) &= \frac{\tilde{f}(t)}{1-\tilde{F}(t)} \\ &= \frac{\underline{e_0} e^{\mathbf{Q}^* t} \underline{v}_P}{1 - \underline{e_0} \left(e^{\mathbf{Q}^* t} - \mathbf{I}\right)) \left(\mathbf{Q}^*\right)^{-1} \underline{v}_P}, \end{aligned} \tag{3.34}$$

where $\tilde{f}(t) = f_{0P}(t)$ and $\tilde{F}(t) = F_{0P}(t)$.

The associated survival function is

$$S(t) = P(T_P > t) = 1 - \tilde{F}(t) = 1 - \underline{e_0} \left(e^{\mathbf{Q}^* t} - \mathbf{I})\right) \left(\mathbf{Q}^*\right)^{-1} \underline{v}_P. \tag{3.35}$$

3.3.4 Expected rates

As the number of regulatory regions z may be unknown, we now introduce the *expected pseudogenization rate* that depends on some distribution of z.

First, we define a random variable Z taking integer values from $Z_{\min}$ to $Z_{\max}$, and let $p_z = P(Z = z)$ be the probability of having z regulatory regions in the duplicate pair of genes. Also, we let $\underline{\mathbf{p}} = [p_z]_{z=Z_{\min},\ldots,Z_{\max}}$ be the (row) vector recording these probabilities. If data is available, the probabilities p_z can of course be estimated as

$$p_z \approx \frac{x_z}{\sum_{i=Z_{\min}}^{Z_{\max}} x_i}, \tag{3.36}$$

where x_z is the observed number of genes with z regulatory regions [53].

Denoting the pseudogenization rate for fixed z by $h(t; z)$ we define the expected pseudogeinzation rate as

$$H(t) = \sum_{z=Z_{\min}}^{Z_{\max}} p_z h(t; z). \tag{3.37}$$

Analogously, we define expected hazard and cause-specific hazard rates (assuming start in state $i = 0$ and dropping the associated subscript) as

$$\Lambda(t) = \sum_{z=Z_{\min}}^{Z_{\max}} p_z \lambda(t; z), \tag{3.38}$$

and

$$\Lambda_j(t) = \sum_{z=Z_{\min}}^{Z_{\max}} p_z \lambda_j(t; z), \tag{3.39}$$

respectively.

Since z is not involved in the limit expressions for h, λ, or λ_j it is immediately clear (applying Equation (3.18) and Corollaries 4 and 5) that

$$\lim_{t\to\infty} \Lambda(t) = 2u_r + u_c, \tag{3.40}$$

$$\lim_{t\to\infty} \Lambda_j(t) = \begin{cases} ur + u_c & \text{for } j = P \\ ur & \text{for } j = S, \end{cases} \tag{3.41}$$

$$\lim_{t\to\infty} H(t) = 0. \tag{3.42}$$

3.4 Hughes and Liberles approximation to the hazard rate

The literature that deals with the subfunctionalization process from a mathematical modelling perspective is limited. The early (and widely cited) work of Force [42] and Force and Lynch [82; 83] introduced the assumption of Poisson rates of mutation in the regulatory and coding regions, and derived some of the measures we covered in Section 3.2. Hughes and Liberles [53] were responsible for perhaps the most detailed analysis since the work of Force and Lynch [42; 82; 83]. In particular they presented an approximation to what they call the pseudogenization hazard rate, but until our recent contribution [109] no mathematical model for the overall process was explicitly set out.

Using their approximate pseudgeonization hazard rate, Hughes and Liberles [53] characterised the subfunctionalization process as having a broadly concave decreasing hazard rate. They contrasted this to a convex decreasing hazard rate associated with neofunctionalization (derived by similar approximation), which they argued was more inline with empirical reality.

Subsequently (e.g. Konrad et al. [66], Tuefel et al. [117]) the analysis of hazard rates by Hughes and Liberles [53] has been used as a reference to define phenomenological approximations to the rate of pseudogenization for gene duplicates evolving under subfunctionalization. The approximations are phenomenological in the sense that functions are chosen to produce the shape properties discussed by Hughes and Liberles [53] without further analysis of the mechanics of the biological model.

We contend in [109] that this focus on hazard rates is slightly misplaced. Since the datasets which are ultimately analysed only detect pseudogenization, and not subfunctionalization, it should be the function in Equation (3.34), rather than the hazard rate (or approximations to it), which is fit to data, and hence which is of principle interest.

This difference is partly semantic, since the function Hughes and Liberles [53] wrote is in practice an approximation to a what we call the cause-specific pseudogenization rate for most of its definition, before switching to an approximation to what we call the pseudogenization rate (or pseudogenization modified-cause-specific hazard rate).

To explicate, Hughes and Liberles [53] applied the following approximation (using the notation introduced in Section 3.2):

$$\lambda_t^z \approx \frac{P_i^z}{E(\Delta T_i)} \quad \text{for } t_{i-1} \leqslant t < t_i, \tag{3.43}$$

where the fixed points t_i are evaluated using

$$t_0 = 0 \quad \text{and} \quad t_i = t_{i-1} + E(\Delta T_i) \quad \text{for } 1 \leqslant i \leqslant z. \tag{3.44}$$

That is, the (approximating) assumption was made that the hazard rates are piecewise constant within such specified time intervals $[t_{i-1}, t_i]$. For $t > t_z$, λ_t^z was assumed to be 0. They wrote,

$$\lambda_t^z = \begin{cases} 2u_c & \text{for } 0 \leqslant t < t_1 \\ u_c & \text{for } t_1 \leqslant t < t_{z-1} \\ u_c + u_r & \text{for } t_{z-1} \leqslant t < t_z \\ 0 & \text{for } t \geqslant t_z. \end{cases} \tag{3.45}$$

No weight is given to the possibility that subfunctionalization has occurred for $t < t_z$, and for $t \geqslant t_z$ no weight is given to the possibility that it has not occurred. While it is true that by the time z mutations have occurred either subfunctionalization or pseudogenization must have occurred, this approximation implicitly assumes that subfunctionalization occurs at the expected time of the z^{th} mutation $t = t_z$ exactly. If the rate remained at $u_c + u_r$ for all $t > t_{z-1}$, so that

$$\lambda_t^z = \begin{cases} 2u_c & \text{for } 0 \leqslant t < t_1 \\ u_c & \text{for } t_1 \leqslant t < t_{z-1} \\ u_c + u_r & \text{for } t \geqslant t_{z-1}, \end{cases} \tag{3.46}$$

this would be a reasonable approximation to the cause-specific hazard rate. However switching to 0 after time t_z indicates that the intent was to make an approximation to something akin to our pseudogenization rate instead. Hughes and Liberles [53] plotted the average of this approximation averaged over a range of z to attempt to infer the shape of the true hazard function — Figure 3.2 shows our recreation of such a plot, similar to Fig. 7 in their paper. Although some of the other examples they

looked at ended in short periods of convex decrease, they nonetheless characterised the pseudogenization rate of the subfunctionalization model as a 'broadly concave decreasing' function. This characterization is at odds with the predictions of our model, which we discuss further in Section 3.5.

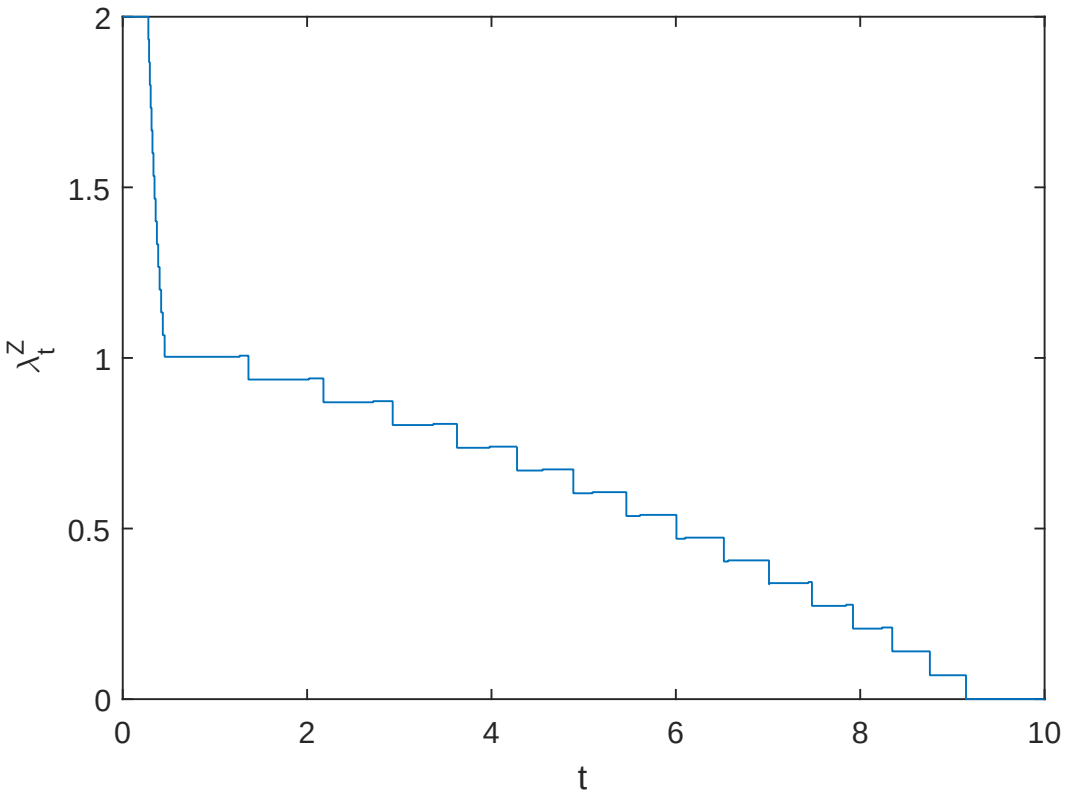

Figure 3.2: A partial recreation of Hughes and Liberles [53] 'Fig. 7' showing their approximation to the mean rate of pseudogenization λ_t^Z with $Z \sim \text{Uni}(2, 16)$, $u_c = 1, u_r = 0.5$.

The focus on shape properties of this approximation is important since this characterization has been used as the basis for various continuous-phenomenological approximations (e.g. in [66; 117; 116]). Different parameterizations of these models are intended to represent the different biological models in the literature, based on the shape properties associated with the parameters. Subfunctionalization is associated with parameters which yield a concave decreasing function, while neofunctionalization is associated with a sigmoid shape ending in a period of convex decrease. We contend that subfunctionalization actually produces behaviour very similar to that which is usually associated with neofunctionalization, and as such this approach to distinguishing between the two biological models may be flawed.

3.5 Shape properties of the pseudogenization rate function

In this section we investigate the shape properties of the pseudogenization rate function $h(t)$ (Equation (3.34)), and compare it to the characterisation of Hughes and Liberles [53]. We show through intuitive argument and numerical confirmation that for the parameter sets of primary biological interest (those with $u_r < u_c$) our model predicts a sigmoid (S-shaped) or exponential-like pseudogenization rate, although richer behaviours are possible when $u_r > u_c$. While the model can yield a concave decreasing hazard rate, as suggested by Hughes and Liberles [53], we show below that this is not necessarily characteristic of the pseudogenization rate implied by our model for subfunctionalization. We find the critical value at which the function's behaviour shifts from being obviously sigmoid (with a period of concave decline) to apparently exponential (i.e. convex decreasing for all t), which we illustrate in Figure 3.5d.

First, we note that $h(t)$ clearly cannot be strictly concave decreasing, since it is bounded below by 0, and a strictly concave decreasing function cannot be bounded below. This is of course not what Hughes and Liberles [53] meant when they characterised the function as 'broadly concave decreasing'. In fact for some of the distributions of Z investigated in [53] the approximation indicated a period of convex decline for large t, but it was always preceded by a long period of concave decline, hence 'broadly concave decreasing'. We find that, while a broadly concave decreasing rate may be sufficient to identify subfunctionalization (assuming that other models do not reproduce the behaviour), it is certainly not necessitated by the model — thus the conclusion that subfunctionalization is a rare mode of duplicate evolution, based on the shape of its associated hazard function, is not supported.

An alternative expression for the pseudogenization rate

Throughout this section we consider $h(t)$ as an average of the rate of transition from state i to P weighted by the probability that the process is in state i at time t (conditional on starting from state 0 and not being absorbed into P).

Proposition 20.

For all $t \geqslant 0$,

$$h(t) = 2u_c p_{\{0\}}(t) + u_c p_{\{1,\ldots,z-2\}}(t) + (u_c + u_r) p_{\{z-1\}}(t) + 0 p_{\{S\}}(t), \tag{3.47}$$

where, for any set $\mathcal{J} \subset \mathcal{S}$

$$p_{\mathcal{J}}(t) = P(X(t) \in \mathcal{J} \mid X(0) = 0, X(t) \neq P). \tag{3.48}$$

Proof.

From Equations (3.24) and (3.26) we have,

$$\begin{aligned} h(t) &= \frac{\tilde{f}(t)}{1 - \tilde{F}(t)} \\ &= \frac{\underline{e_0} e^{\mathbf{Q}^* t} \underline{v}_P}{1 - \tilde{F}(t)} \\ &= \frac{\sum_{i \in \{0,1,\ldots,z-1\}} P_{0i}(t) v_{iP}}{1 - \tilde{F}(t)}, \end{aligned}$$

and so, noting that

$$1 - \tilde{F}(t) = P(X(t) \neq P \mid X(0) = 0), \tag{3.49}$$

we get

$$\begin{aligned} h(t) &= \frac{1}{P(X(t) \neq P \mid X(0) = 0)} \sum_{i \in \{0,1,\ldots,z-1\}} P_{0i}(t) v_{iP} \\ &= \frac{1}{P(X(t) \neq P \mid X(0) = 0)} \sum_{i \in \{0,1,\ldots,z-1\}} P(X(t) = i \mid X(0) = 0)) v_{iP}. \end{aligned}$$

By the fact that for $i \neq P$ events $\{X(t) = i\}$ and $\{(X(t) = i) \cap (X(t) \neq P)\}$ are equivalent, and the definition of conditional probability, we get

$$\begin{aligned} h(t) &= \sum_{i \in \{0,1,\ldots,z-1\}} \frac{P(X(t) = i, X(t) \neq P \mid X(0) = 0)}{P(X(t) \neq P \mid X(0) = 0)} v_{iP} \\ &= \sum_{i \in \{0,1,\ldots,z-1\}} \frac{P(X(t) = i, X(t) \neq P, X(0) = 0) P(X(0) = 0)}{P(X(t) \neq P, X(0) = 0) P(X(0) = 0)} v_{iP} \\ &= \sum_{i \in \{0,1,\ldots,z-1\}} \frac{P(X(t) = i, X(t) \neq P, X(0) = 0)}{P(X(t) \neq P, X(0) = 0)} v_{iP} \\ &= \sum_{i \in \{0,1,\ldots,z-1\}} P(X(t) = i \mid X(0) = 0, X(t) \neq P) v_{iP} \\ &= \sum_{i \in \{0,1,\ldots,z-1\}} p_{\{i\}}(t) v_{iP} \\ &= \sum_{i \in \{0,1,\ldots,z-1\}} q_{iP} p_{\{i\}}(t). \end{aligned} \tag{3.50}$$

Now,

$$q_{iP} = \begin{cases} 2u_c & \text{for } i = 0 \\ u_c & \text{for } i = 1, ..., z-2 \\ u_c + u_r & \text{for } i = z-1, \end{cases}$$

so we get,

$$h(t) = 2u_c p_{\{0\}}(t) + u_c \sum_{i \in \{1,..,z-2\}} p_{\{i\}}(t) + (u_c + u_r) p_{\{z-1\}}(t).$$

By the addition rule for disjoint events we have,

$$\sum_{i \in \{1,..,z-2\}} p_{\{i\}}(t) = p_{\{1,...,z-2\}}(t),$$

so,

$$\begin{aligned} h(t) &= 2u_c p_{\{0\}}(t) + u_c p_{\{1,...,z-2\}}(t) + (u_c + u_r) p_{\{z-1\}}(t) \\ &= 2u_c p_{\{0\}}(t) + u_c p_{\{1,...,z-2\}}(t) + (u_c + u_r) p_{\{z-1\}}(t) + 0 p_{\{S\}}(t). \end{aligned}$$

□

The following more general result is attained in analogous manner, stopping at Equation (3.50).

Corollary 6.

For an absorbing CTMC with state space $\mathcal{S} = \mathcal{S}^* \cup \mathcal{A}$, for all $i \in 1, ..., m, j \in \mathcal{A}$ we have,

$$\lambda_i^j(t) = \sum_{k \in \mathcal{S}^*} q_{kj} p_k^{ij}(t), \tag{3.51}$$

where $p_k^{ij}(t) = P(X(t) = k \mid X(0) = i, X(t) \neq j)$.

Period of increase

Considering the structure of the Markov chain (with generator defined by Equation (3.2)), we can see that each transient state is visited at most once, and that they are visited in order. Since the process visits each transient state at most once, followed by an exponential time before leaving, we can conclude the following:

- $p_{\{0\}}(t)$ is decreasing for all t from an initial value of 1 towards 0;

- $p_{\{S\}}(t)$ increases for all t from an initial value of 0 towards 1;
- $p_{\{1,...,z-2\}}(t)$ is initially increasing (from 0), but will turn around (once) and approach 0; and
- $p_{\{z-1\}}(t)$ is initially increasing (from 0), but will turn around (once) and approach 0.

By the structure of the Markov chain and the law of total probability, any increase in $p_{\{i\}}(t)$ must be balanced by a decrease in $p_{\{j:j<i\}}(t)$ (adopting $i < S$ for each $i = 1, ..., z-1$ for notational convenience).

In fact, considered in isolation an increase in $p_{\{i\}}(t)$ is balanced by a decrease specifically in $p_{\{i-1\}}(t)$ for $i = 1, ..., z-1$ (but simultaneous changes in $p_{\{j:j<i-1z\}}(t)$ can counterbalance the decrease in $p_{\{i-1\}}(t)$). We make these observations more precise below, first by considering general sets, and then applying the results to the specific sets of interest.

We denote $\mathcal{S}\backslash\{P\}$ by $\mathcal{S}_P$, and note that $p_{\mathcal{S}_P}(t) = 1$ for all $t \geqslant 0$. Then, by the law of total probability, for any $t \geqslant 0$ and any $\mathcal{J} \subseteq \mathcal{S}_P$, we have

$$p_{\mathcal{J}}(t) = 1 - p_{\mathcal{S}_P \backslash \mathcal{J}}(t). \tag{3.52}$$

Hence, for any $\Delta t > 0$, we have,

$$p_{\mathcal{J}}(t + \Delta t) = p_{\mathcal{S}_P}(t) - p_{\mathcal{S}_P \backslash \mathcal{J}}(t + \Delta t). \tag{3.53}$$

Thus, for any $t, \Delta t > 0$, and any sets $\mathcal{J}, \mathcal{I}, \mathcal{K}$ which partition $\mathcal{S}_P$, if $p_{\mathcal{K}}(t)$ is held constant, we get

$$\begin{aligned} p_{\mathcal{J}}(t + \Delta t) &= p_{\mathcal{S}_P}(t) - p_{\mathcal{S}_P \backslash \mathcal{J}}(t + \Delta t) \\ &= p_{\mathcal{J}}(t) + p_{\mathcal{I}}(t) + p_{\mathcal{K}}(t) - p_{\mathcal{I}}(t + \Delta t) - p_{\mathcal{K}}(t + \Delta t) \\ &= p_{\mathcal{J}}(t) + p_{\mathcal{I}}(t) - p_{\mathcal{I}}(t + \Delta t). \end{aligned} \tag{3.54}$$

Now, if we let $p_{\mathcal{I}}(t) - p_{\mathcal{I}}(t + \Delta t) = \epsilon$ for some $\epsilon \in \mathbb{R}$, we have,

$$p_{\mathcal{J}}(t + \Delta t) = p_{\mathcal{J}}(t) + \epsilon. \tag{3.55}$$

Applying Equation (3.55) to the sets from Equation (3.47), we have

- For any $t, \Delta t > 0$, holding $p_{\{z-1,S\}}(t)$ constant, if, for any $\epsilon \in \mathbb{R}$,

$$p_{\{1,...,z-2\}}(t + \Delta t) = p_{\{1,...,z-2\}}(t) + \epsilon, \tag{3.56}$$

then,

$$p_{\{0\}}(t+\Delta t) = p_{\{0\}}(t) - \epsilon. \tag{3.57}$$

Combined, the effect on h is a decrease of $u_c\epsilon$, i.e.

$$h(t+\Delta t) = h(t) - u_c\epsilon. \tag{3.58}$$

- Similarly if, holding $p_{\{0,S\}}(t)$ constant, we have

$$p_{\{z-1\}}(t+\Delta t) = p_{\{z-1\}}(t) + \epsilon, \tag{3.59}$$

then,

$$p_{\{1,\ldots,z-2\}}(t+\Delta t) = p_{\{1,\ldots,z-2\}}(t) - \epsilon. \tag{3.60}$$

The combined effect on h is an increase of $u_r\epsilon$, i.e.

$$h(t+\Delta t) = h(t) + u_r\epsilon. \tag{3.61}$$

- Also, holding $p_{\{0\}}(t)$ constant, if

$$p_{\{S\}}(t+\Delta t) = p_{\{S\}}(t) + \epsilon, \tag{3.62}$$

then we have

$$p_{\{1,\ldots,z-2\}}(t+\Delta t) + p_{\{z-1\}}(t+\Delta t) = p_{\{1,\ldots,z-2\}}(t) + p_{\{z-1\}}(t) - \epsilon. \tag{3.63}$$

The combined effect on h gives

$$h(t) - (u_r + u_c)\epsilon \leqslant h(t+\Delta t) \leqslant h(t) - u_c\epsilon, \tag{3.64}$$

with the lower bound attained in the case that $p_{\{1,\ldots,z-2\}}(t+\Delta t) = p_{\{1,\ldots,z-2\}}(t)$, and the upper bound attained in the case $p_{\{z-1\}}(t+\Delta t) = p_{\{z-1\}}(t)$.

Furthermore, since transitions into state $z-1$ only occur from $z-2$, and the rate of transition from $z-2$ to S is equal to the rate of transition from $z-2$ to $z-1$ it follows that $p'_{\{z-1\}}(t) \leqslant p'_{\{S\}}(t)$ for all t. If $z > 2$ then there is at least one other transient state with a non-zero transition rate to S, so that $p'_{\{z-1\}}(t) < p'_{\{S\}}(t)$.

Hence for $h(t)$ to increase during any period, the overall (positive) contribution from increasing $p_{\{z-1\}}(t)$ has to exceed the (negative) contribution from the simultaneous increase of $p_{\{S\}}(t)$, and the increase in $p_{\{S\}}(t)$ is guaranteed to be at least as large as the increase in $p_{\{z-1\}}(t)$ at all times. Since the first contribution is directly proportional to u_r, and the second to at least u_c (in absolute value), $u_r > u_c$ is a necessary but not

sufficient condition for there to be any period of increase in $h(t)$. After the period of increase in $p_{\{z-1\}}(t)$, $h(t)$ is strictly decreasing.

More precisely, consider the change in h over a period $[t, t+\Delta t]$, given by

$$h(t+\Delta t) - h(t) = 2u_c\epsilon_1 + u_c\epsilon_2 + (u_c+u_r)\epsilon_3 + 0\epsilon_4, \tag{3.65}$$

where ϵ_i is the change in the probability factor of the i^{th} term in the right hand side of Equation (3.47) over the period $[t, t+\Delta t]$, for example, $\epsilon_1 = p_{\{0\}}(t+\Delta t) - p_{\{0\}}(t)$, and $\epsilon_2 = p_{\{1,..,z-1\}}(t+\Delta t) - p_{\{1,...,z-1\}}(t)$, etc.

Now, we know from the above discussion that that for all $t, \Delta t$, we have $\epsilon_1 < 0$ and $\epsilon_4 > 0$.

First, we consider the case where $\epsilon_2, \epsilon_3 > 0$, (which is always the case during the period $[0, \lim_{\Delta t\to 0} \Delta t]$) then from Equation (3.65) we can write,

$$h(t+\Delta t) - h(t) < 2u_c\epsilon_1 + (u_c+u_r)(\epsilon_2+\epsilon_3). \tag{3.66}$$

From the law of total probability we know that $\epsilon_2 + \epsilon_3 = -\epsilon_1 - \epsilon_4$, thus

$$\begin{aligned} h(t+\Delta t) - h(t) &< 2u_c\epsilon_1 - (u_c+u_r)(\epsilon_1+\epsilon_4) \\ &= u_c(\epsilon_1-\epsilon_4) + u_r(\epsilon_4-\epsilon_1). \end{aligned} \tag{3.67}$$

Since we know that for all $t, \Delta t$, we have $\epsilon_1 < 0$ and $\epsilon_4 > 0$, the first term is necessarily negative, and the second positive, so we have

$$h(t+\Delta t) - h(t) < u_r|\epsilon_1-\epsilon_4| - u_c|\epsilon_1-\epsilon_4|. \tag{3.68}$$

Thus, if $u_r < u_c$, h is decreasing over the period during which $\epsilon_2, \epsilon_3 > 0$.

The other case we need to consider is that of a period during which $\epsilon_2 < 0, \epsilon_3 > 0$. Since $\epsilon_3 = -\epsilon_1 - \epsilon_2 - \epsilon_4$. Such an interval may not exist in practice, but we will assume it does for now (otherwise the above argument is sufficient to show that $h(t)$ is non-increasing). We can rewrite Equation (3.65) as

$$\begin{aligned} h(t+\Delta t) - h(t) &= 2u_c\epsilon_1 + u_c\epsilon_2 + (u_c+u_r)(-\epsilon_1-\epsilon_2-\epsilon_4) \\ &= u_c(\epsilon_1-\epsilon_4) + u_r(-\epsilon_1-\epsilon_2-\epsilon_4). \end{aligned} \tag{3.69}$$

Since $\epsilon_4 \geqslant \epsilon_3$ for all $t, \Delta t$, we have

$$h(t+\Delta t) - h(t) \leqslant u_c(\epsilon_1-\epsilon_4) + u_r\epsilon_4, \tag{3.70}$$

and since $\epsilon_1 < 0$ for all $t, \Delta t$, we have

$$h(t+\Delta t) - h(t) \leqslant -u_c\epsilon_4 + u_r\epsilon_4, \tag{3.71}$$

thus, for h to increase over any period during which $\epsilon_2 < 0, \epsilon_3 > 0$ we also require $u_r > u_c$. Clearly if $\epsilon_3 < 0$ h is not increasing, by the itemised argument above. Therefore, for h to exhibit any period of increase, it is necessary that $u_r > u_c$.

We would expect the relative size of u_r which yields a period of increase in $h(t)$, to increase with z (since increasing z means the process has to transition through more transient states before reaching $z-1$). During our numerical analysis we found $u_r = 1.01u_c$ will produce a (very small) period of increase in $h(t)$ for $z = 2$, but $z = 3$ requires $u_r > 4.0u_c$, and $z = 4$ requires $u_r > 6.8u_c$. Based on the physical intuition, we suspect that this trend continues, and no counterexample was encountered during our numerical analysis.

Example 3.5.1 ($h(t)$ with an initial period of increase)**.**

In this example, we examine the shape of the pseudogenization rate $h(t)$ for $u_c = 1$, $u_r = 1.9$, $z = 2$. In this case, there is a clear period of increase for small t, as shown in Figure 3.3.

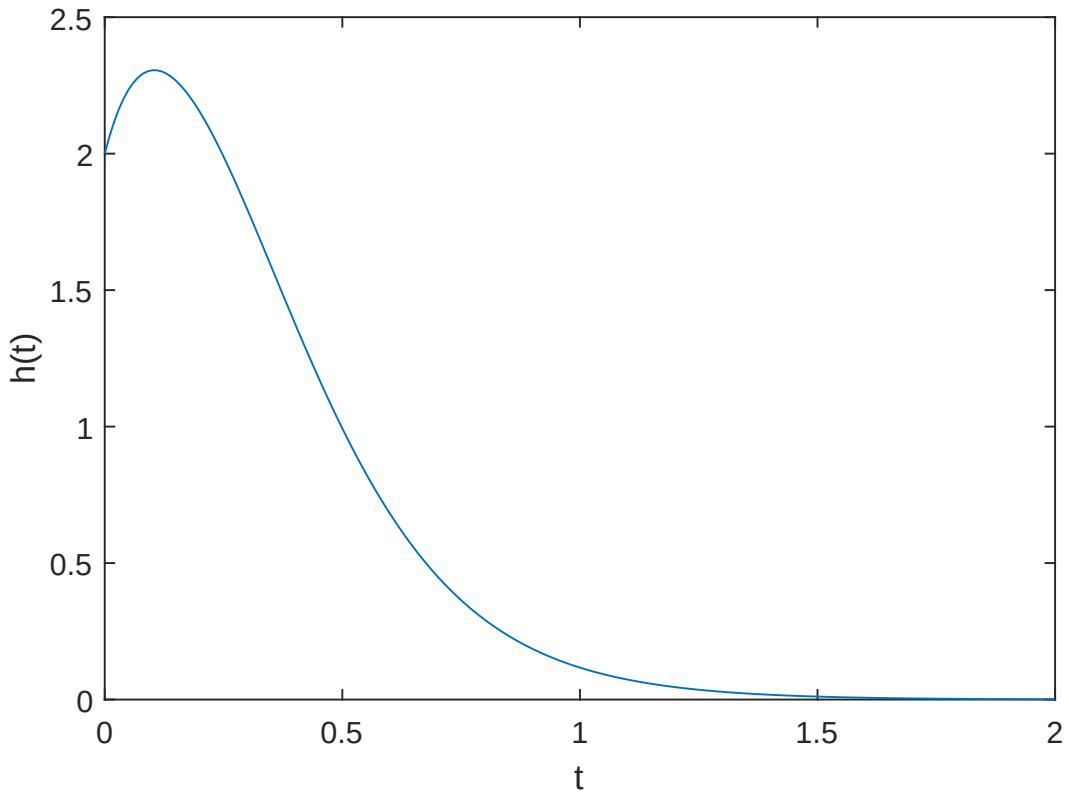

Figure 3.3: Pseudogenization rate $h(t)$ for $u_c = 1, u_r = 1.9, z = 2$ — a short period of increase is seen before the function begins to decrease towards its limit of 0.

Combining all of this, we have that $h(t)$ can only increase during a single period in

which $p_{\{z-1\}}(t)$ is increasing, and will only do so only if u_r is at least as large as u_c; for most z only if $u_r >> u_c$, which is generally not considered to be biologically realistic. At all other times $h(t)$ is decreasing, and is convex at least for large t. These observations are all confirmed by numerical examination — we performed a grid search of the parameter space (with $u_c = 1$ fixed as per Remark 4) to identify distinct qualitative behaviours of $h(t)$, the results of which were in agreement with this analysis.

Points of inflection

The results of our grid search of the parameter space suggest that when u_r is small, $h(t)$ has a single point of inflection, while as u_r is increased, there is no point of inflection until some threshold at which two points of inflection appear (which occurs at some $u_r > u_c$).

Our understanding of this phenomena is that when u_r is sufficiently small, $p_{\{0\}}(t)$ initially declines slowly, and there is a delay before $p_{\{S\}}(t)$ becomes significant, resulting in an initially concave decline in $h(t)$. Since $h(t)$ must be convex declining for large t, a point of inflection must therefore occur. Based on this reasoning and the numerical results, it appears that there is at most one point of inflection when $u_r < u_c$, and we refer to it as 'the' point of inflection, or 'the' change in concavity when we examine this case further below.

On the other hand, when u_r is large, $h(t)$ is initially convex decreasing, but during the period in which $p_{\{z-1\}}(t)$ is increasing, its contribution can slow the decline of $h(t)$ sufficiently to produce a change in concavity (whether or not it also leads to an overall increase in $h(t)$). Again, since $h(t)$ must ultimately be convex declining, a second point of inflection must occur.

Example 3.5.2 ($h(t)$ with two points of inflection)**.**

In this example, we examine the shape of the pseudogenization rate $h(t)$ for $u_c = 1$, $u_r = 3$, $z = 3$. We can see clearly that there are two points of inflection, and, despite u_r being somewhat larger than in Example 3.5.1, incrementing z ensures that there is no period of increase. This example is plotted in Figure 3.4.

A point of inflection will occur when $h''(t) = 0$; applying the quotient rule and differentiating Equation (3.24) with respect to t twice, we get (omitting the arguments of the functions on the right hand side for brevity),

$$h''(t) = \frac{[f''(1-F) - (1-F)''f]\,(1-F) - 2(1-F)'\,[f'(1-F) - (1-F)'f]}{(1-F)^3}. \quad (3.72)$$

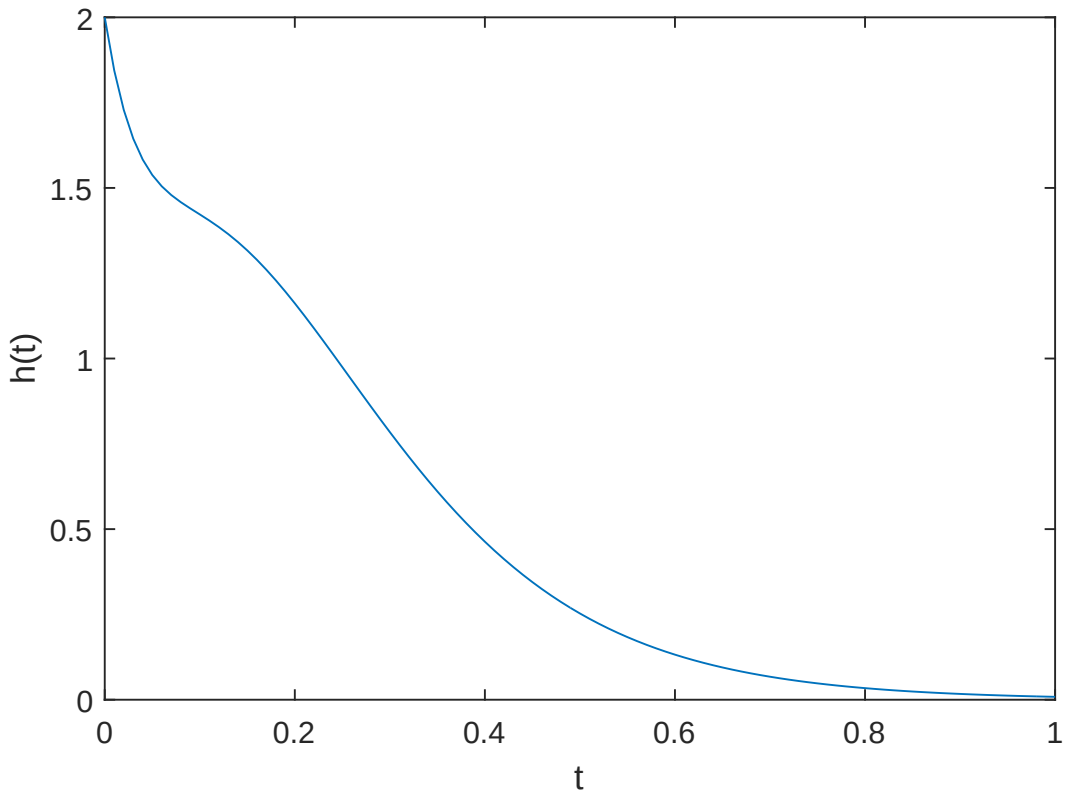

Figure 3.4: Pseudogenization rate $h(t)$ for $u_c = 1, u_r = 3, z = 3$, showing an initial period of convex decrease, followed by two changes in concavity.

Recalling that f and F are the density and cumulative distribution functions associated with time to pseudogenization, we have $f = F'$, and hence,

$$\begin{aligned} h''(t) &= \frac{\left[F'''(1-F) - (1-F)''F'\right](1-F) - 2(1-F)'\left[F''(1-F) - (1-F)'F'\right]}{(1-F)^3} \\ &= \frac{\left[F'''(1-F) + F''F'\right](1-F) + 2F'\left[F''(1-F) + F'F'\right]}{(1-F)^3} \\ &= \frac{F'''(1-F)^2 + 3F''F'(1-F) + 2(F')^3}{(1-F)^3}. \end{aligned} \tag{3.73}$$

Thus, the change in concavity will occur when

$$F'''(1-F)^2 + 3F''F'(1-F) + 2(F')^3 = 0. \tag{3.74}$$

No further simplification was achievable, but this form is suitable for finding the point of inflection numerically, noting that for $n \in \mathbb{Z}^+$,

$$F^{(n)} = \underline{e}_0 e^{\boldsymbol{Q}^* t}(Q^*)^{n-1}\underline{v}_P, \tag{3.75}$$

where $F^{(n)}$ denotes F differentiated with respect to t n times.

Case with $u_r < u_c$

For the case of primary biological interest, with $u_r < u_c$, $h(t)$ is strictly decreasing, with at most one point of inflection as per the above arguments. In this case, the most important qualitative feature in terms of biological interpretation is the presence or absence of the change in concavity, determining whether the characteristic period of concave decrease is observed.

The presence of a change in concavity is well illustrated graphically by extending the domain of $h(t)$ from $\mathbb{R}^+ \cup \{0\}$ to $\mathbb{R}$, and observing whether the point of inflection occurs for $t \geqslant 0$, or $t < 0$. The expression for the function (which is well behaved over $\mathbb{R}$) is unchanged from Equation (3.34), but the physical interpretation for $h(t)$ as the pseudogenization rate at time t after duplication is restricted to the domain $\mathbb{R}^+ \cup \{0\}$. The presence or absence of a point of inflection in $\mathbb{R}^+ \cup \{0\}$ leads to two distinct biological predictions for the behaviour of the process.

- The duplication can be characterised by a rapid fixation of the duplicate copies by subfunctionalization so that pseudogenization is most likely to occur immediately, or not-at-all. This is associated with the absence of the point of inflection, and an immediate convex decline in the pseudogenization rate function, which quickly approaches its asymptote at zero, leaving a relatively short window during which the pseudogenization rate is far from zero. This is totally at-odds with the characterization of Hughes and Liberles [53].
- Alternatively, it can be characterised by a slower fixation process — subfunctionalization is unlikely to occur for some time, and hence there is a significant period during which at least one copy is vulnerable to pseudogenization. This is associated with the presence of an initially concave decreasing rate of pseudogenization and a (long or short) initial period during which the rate is relatively flat. This is more consistent with the characterization of Hughes and Liberles [53], although the period of concave decrease can be arbitrarily short, and is always followed by a point of inflection and a long period of convex decrease.

The line between these two behaviours is not completely clear cut, since having a point of inflection occur at or near $t = 0$ is not much of a distinction from having no point of inflection at all, as can be seen by comparing Figures 3.5c and 3.5b. Nonetheless, we treat these two behaviours as distinct for this discussion.

Remark 6.

In the extended domain $\mathbb{R}$ it is possible for $h(t)$ to have additional points of inflection

for some parameter choices — obviously the physical intuition offers no insight into the function's behaviour for negative values of t.

Definition 56 (Critical rate ratio γ^z_{crit})**.**

We define γ^z_{crit} to be the ratio u_r/u_c at which the change in concavity occurs precisely when t is zero, i.e the ratio u_r/u_c such that Equation (3.74) is satisfied only for $t = 0$.

When $0 < \gamma = u_r/u_c < \gamma^z_{\text{crit}}$ the concavity of $h(t)$ will change at some $t^* > 0$, (with t^* increasing as γ decreases) and we see the behaviour where an initially slowly-declining pseudogenization rate decreases more and more quickly before slowing back down as it approaches zero. Otherwise, the change in concavity does not exist in $\mathbb{R}^+ \cup \{0\}$, and in this case the pseudogenization rate begins its rapid decline immediately, with the rate of decline slowing from its initially-high value at all times. In the gene duplication context, this behaviour is typically thought to be characteristic of neo-, and not subfunctionalization, as described by Hughes and Liberles [53]. Figure 3.6 shows the values of γ^z_{crit} for various values of z.

Example 3.5.3 ($\gamma < \gamma^z_{\text{crit}}$)**.**

In this example, we examine the shape of the pseudogenization rate for $z = 12$, and $\gamma = u_r/u_c = 0.005 < \gamma^{12}_{crit} = 0.0714$. In this case, the sigmoidal shape of the rate function is quite apparent. A change in concavity occurs at $t > 0$, and the rate is relatively flat near $t = 0$, as shown in Figure 3.5a.

Example 3.5.4 ($\gamma = \gamma^z_{\text{crit}}$)**.**

For this example, we examine the shape of the pseudogenization rate for $z = 12$ and $\gamma = u_r/u_c = 0.0714 = \gamma^{12}_{crit}$, shown in Figure 3.5c. Here, considered as a function over $\mathcal{R}$, there is a point of inflection at $t = 0$, thus in the domain $\mathcal{R}^+ \cup \{0\}$ we see little evidence of the sigmoidal shape of the rate function, which is qualitatively similar to an exponential decay.

Example 3.5.5 ($\gamma > \gamma^z_{\text{crit}}$)**.**

In this example, we examine the shape of the pseudogenization rate for $z = 12$ and $\gamma = u_r/u_c = 0.2 > \gamma^{12}_{crit} = 0.0714$. In this case, there is no point of inflection in the domain $\mathbb{R}^+ \cup \{0\}$, and the shape of the function is similar to that of an exponential decay. Figure 3.5c shows the rate function for this example.

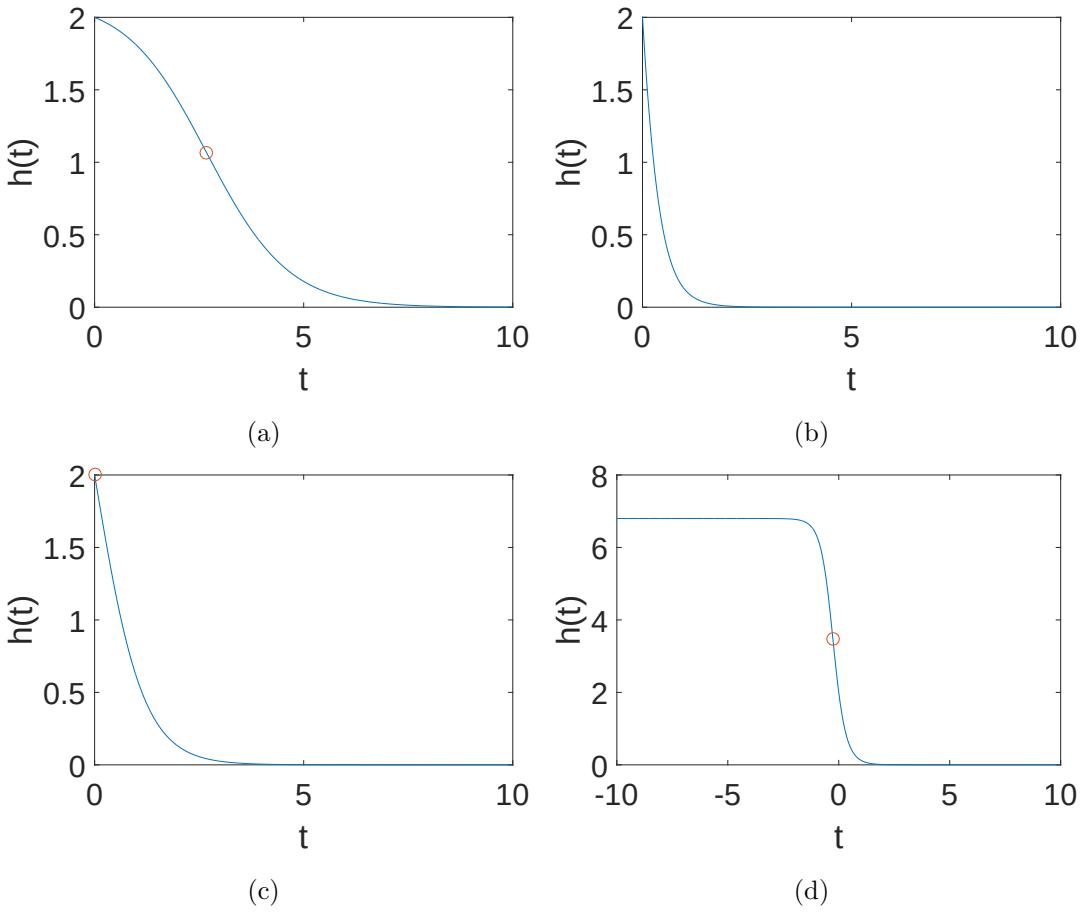

Figure 3.5: Pseudogenization rate $h(t)$ for $z = 12$ and γ less than (a), greater than (b) and equal to (c) γ_{crit}. Panel (d) shows the overall shape of $h(t)$ for the extended domain with negative values of t included. The point of inflection (when visible) is marked with a red circle.

3.6 Comparison of pseudogenization rate to existing phenomenological approximations

We now compare the pseudogenization rate derived in (3.34) to the phenomenological approximations of Konrad et al. [66] and Tuefel et al. [117]. As mentioned in Section 3.4, the characterization derived from Hughes and Liberles' [53] piece-wise constant approximation to the pseudgogenization rate associated with the sub- and neofunctionalization has been used to inform subsequent smooth phenomenological models. The models employed by Konrad et al. [66] and Tuefel et al. [117] are two such models; they are thought to capture the behaviour of the different biological processes under different parameters, with subfunctionalization thought to correspond to

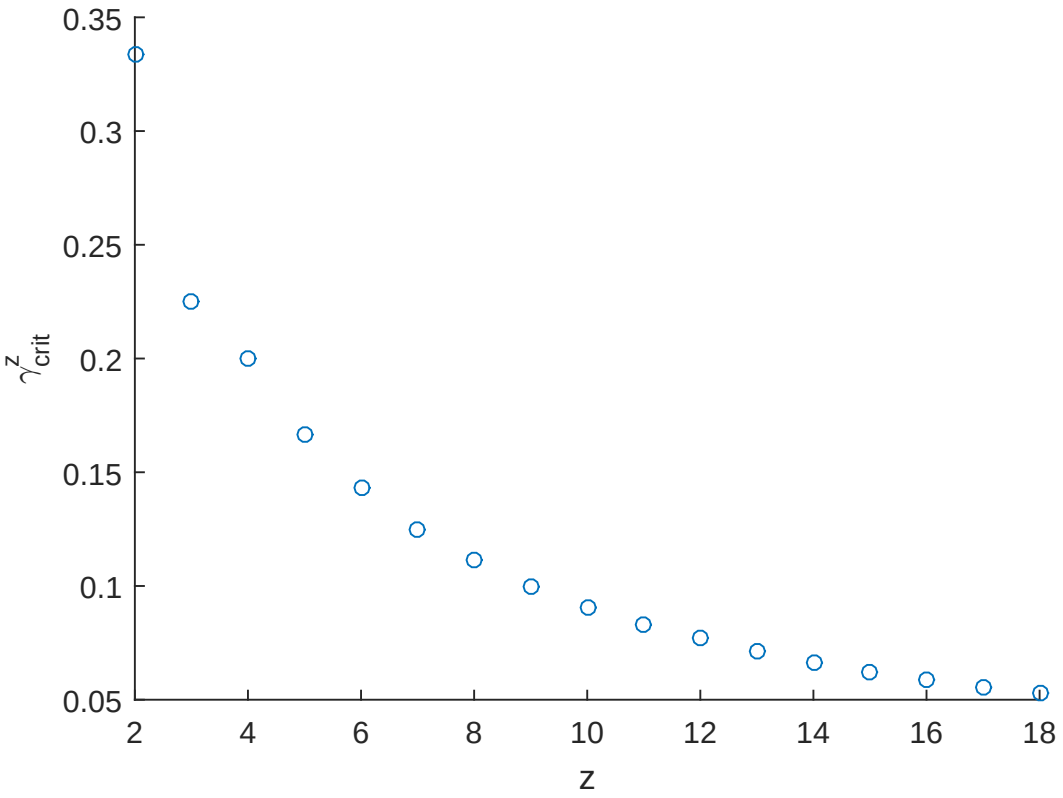

Figure 3.6: Critical values γ^z_{crit} for various values of z. When $u_r/u_c \leqslant \gamma^z_{\text{crit}}$ there will be a change in concavity for the pseudogenization rate function (with domain $\mathbb{R} \cup \{0\}$).

a broadly concave decreasing hazard rate.

Here we analyze the parameter space of both of these phenomenological rate functions and show that more of the parameter space of both of models is inline with the predictions of subfunctionalization than previously thought. We find that both of the approximations have a good qualitative correspondence to the pseuodgenization rate (3.34) derived here, with the approximation in Teufel et al. [117] being in particularly good agreement with our rate function. We then derive some results to select appropriate parameters for our model to reproduce the behaviour for the model (due to Tuefel et al. [117]).

3.6.1 Approximation in Konrad et al. [66]

First, consider the approximation in Konrad et al. [66], which we denote $h_K(t)$, given by

$$h(t) \approx h_K(t) = fe^{-bt^c} + d. \tag{3.76}$$

The first and second derivative of $h_K(t)$ are given by

$$h'_K(t) = -fbct^{c-1}e^{-bt^c} \tag{3.77}$$

$$\begin{aligned} h''_K(t) &= f\left(b^2c^2t^{2(c-1)}e^{-bt^c} - b(c-1)ct^{c-2}e^{-bt^c}\right) \\ &= bct^{c-2}e^{-bt^c}\left(c(bt^c-1)+1\right). \end{aligned} \tag{3.78}$$

Below, we show that the qualitative behaviour of the pseudogenization rate $h(t)$ defined in (3.34) can be reproduced using a function of the form (3.76). We find sets of parameters f, b, c and d that correspond well to the qualitative behaviour of h in each of the three cases $\gamma < \gamma_{\text{crit}}$, $\gamma = \gamma_{\text{crit}}$ and $\gamma > \gamma_{\text{crit}}$.

We are primarily interested in parameter c, since c is a shape parameter, and it is the choice of c that Konrad et al. [66] use to distinguish between sub- and neofunctionalization. Before we discuss parameter c, we note that since $\lim_{t\to\infty} e^{-t} = 0$, then $\lim_{t\to\infty} h_K(t) = d$. The corresponding limit for the pseudogenization rate is $\lim_{t\to\infty} h_P(t) = 0$, and so we require $d = 0$ if we wish to match the long-run behaviour of the approximation $h_K(t)$ to that of $h_P(t)$. Also, since f is a scale parameter, with $h_K(0) = f + d$, choosing $f = 2u_c = h(0)$ and $d = 0$ will match the initial and limit values of the two functions $h(t)$ and $h_K(t)$.

$c = 1$ corresponds well to $\gamma \geqslant \gamma_{\text{crit}}$

When $c = 1$, $h_K(t)$ is exponential. Recalling Examples 3.5.4 and 3.5.5, we note that $h_P(t)$ behaves like an exponential when $\gamma \geqslant \gamma_{\text{crit}}$, so we have good qualitative correspondence between $h_K(t)$ and $h(t)$ when $c = 1$ and $\gamma \geqslant \gamma_{\text{crit}}$.

Figure 3.7 shows a plot of $h_K(t)$ with $c = 1$.

$c > 1$ corresponds reasonably well to $\gamma < \gamma_{\text{crit}}$

When $c > 1$, $h'_K(t)$ given in Equation (3.77) has a single root at $t = 0$, while $h''_K(t)$, given in Equation (3.78) has two roots, one at $t = 0$ and one when $bt^c - 1 = 1/c$, which occurs for some $t > 0$. This means that there is a change of concavity at some $t > 0$, before a flattening out as $t \to 0^+$. This is qualitatively similar to $h(t)$ for $\gamma < \gamma_{\text{crit}}$. This is the parameter set that Konrad et al. [66] intended to correspond to the subfunctionalization process, based on the intuition that there is an observable waiting time for an initial change to allow for subsequent subfunctionalization.

However, for any realistic set of parameters (the exception being $u_r = 0$, in which case the model does not describe the biological process of subfunctionalization) we have

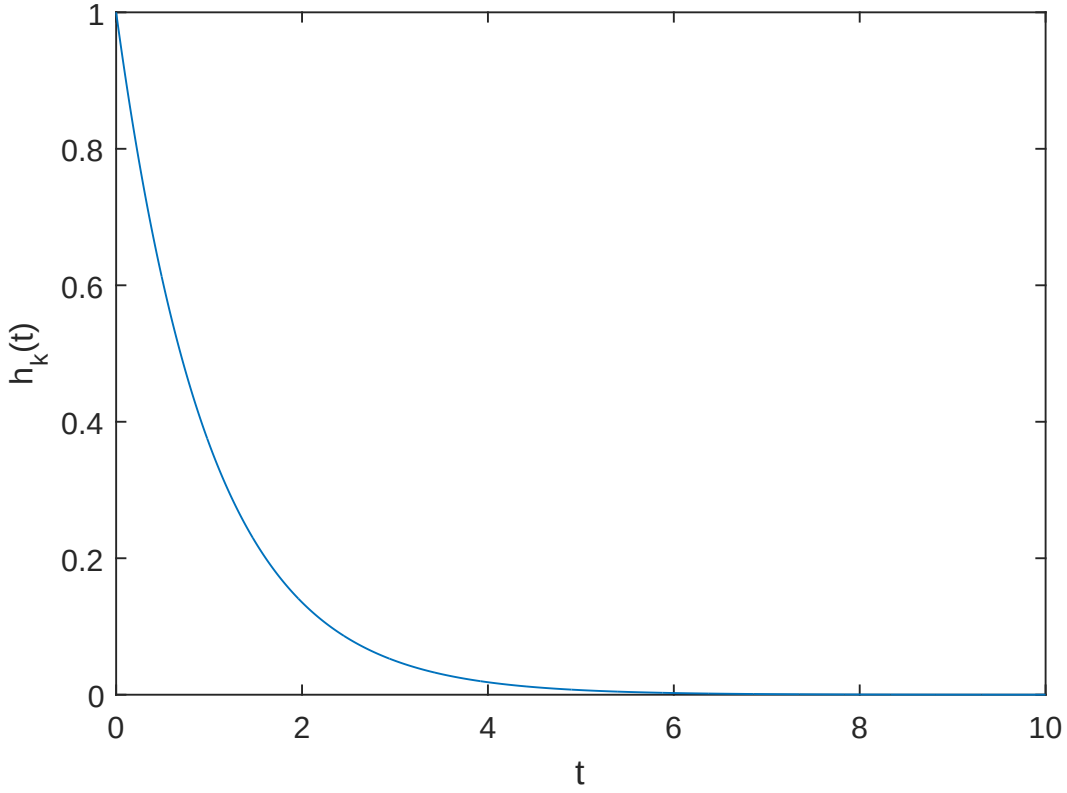

Figure 3.7: The approximation in Konrad et al. [66] with shape parameter $c = 1$. The remaining parameters were $f = b = 1$, $d = 0$.

$h'(t) \neq 0$ for any finite t. As a result, the flattening out behaviour around $t = 0$ is never quite matched by the pseudogenization rate function, unless u_r is so large that that $h(0) \approx \lim_{t\to-\infty} h(t)$ in the extended domain. This is highly unrealistic, besides which the rest of the behaviour of the two functions is very dissimilar in this case.

Figure 3.8 shows $h_K(t)$ for $c = 3 > 1$.

$0 < c < 1$ corresponds somewhat well to $\gamma > \gamma_{\mathbf{crit}}$

The case with $c < 1$ is the most distinct from the subfunctionalization model, with the derivative $h'_K(t) \to -\infty$ as $t \to 0^+$. It is is qualitatively somewhat similar to the function $h(t)$ when $\gamma > \gamma_{\text{crit}}$. However, $h_K(t)$ gives a relatively faster decline in the rate for small t, and a slower decline for large t than is achievable with any parameterization of $h(t)$. This is the part of the parameter space which Konrad et al. [66] use to model neofunctionalization.

Although the agreement between the predictions of the subfunctionalization and $h_K(t)$

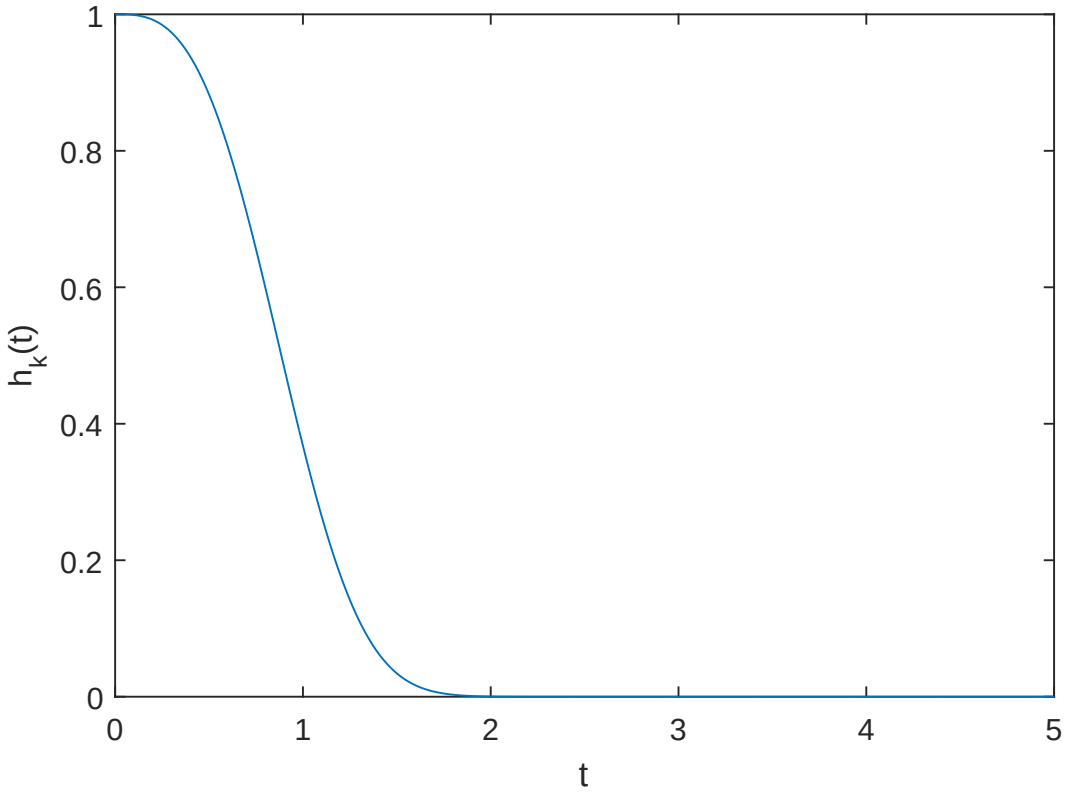

Figure 3.8: The approximation in Konrad et al. [66] with shape parameter $c = 3$. The remaining parameters were $f = b = 1$, $d = 0$.

for $0 < c < 1$ is not as good as it is for the rest of the parameter space, it is significantly better than was previously thought, and we are not convinced that this parameterization can be uniquely associated to neofunctionalization. Further work modelling neofunctionalization will illuminate this, but for now our intuition is that the two processes will be difficult to distinguish by their associated pseudogenization rate functions.

Figure 3.9 shows $h_K(t)$ for $0 < c = 0.5 < 1$.

The remaining cases ($c \leqslant 0$) were not considered biologically realistic by Konrad et al. [66], and are not in agreement with the predictions of the subfunctionalization model. $c = 0$ has no dependence on time, and $c < 0$ would give a strictly increasing rate function.

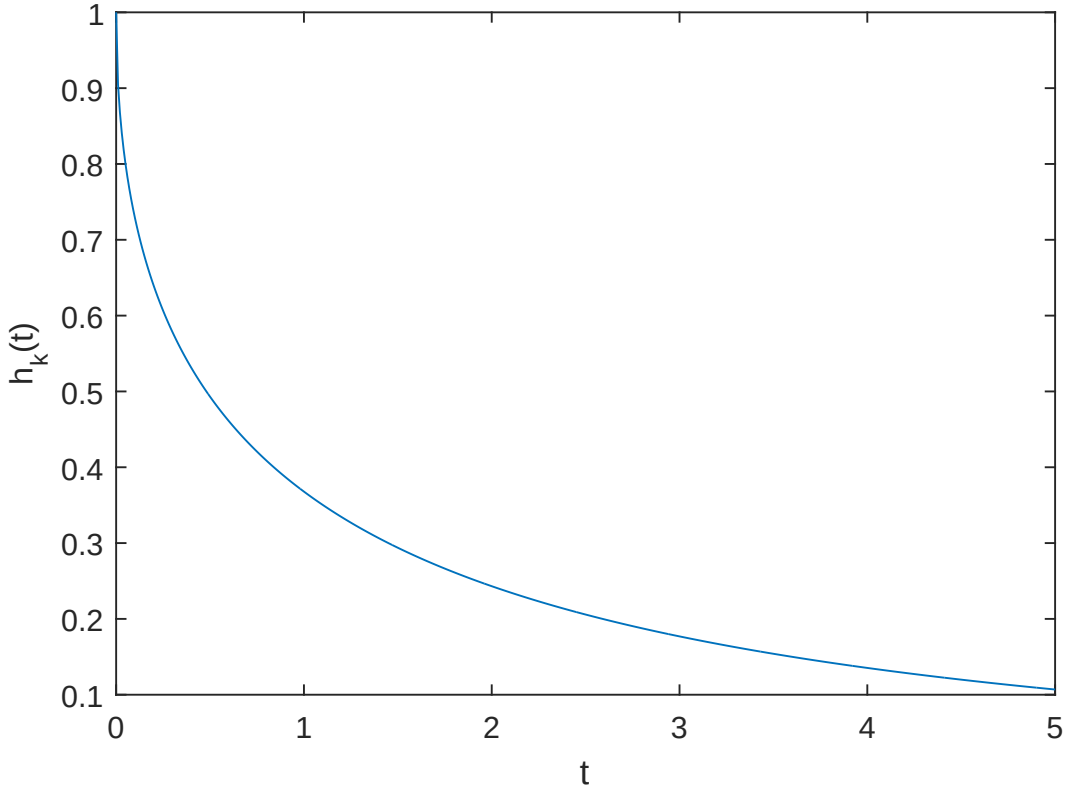

Figure 3.9: The approximation in Konrad et al. [66] with shape parameter $c = 0.5$. The remaining parameters were $f = b = 1$, $d = 0$.

3.6.2 Approximation in Tuefel et al. [117]

We now consider the approximation in Tuefel et al. [117] which we denote $h_T(t)$, given by

$$h(t) \approx h_T(t) = u + \frac{be^{a-t}}{1 + e^{a-t}}. \tag{3.79}$$

The first and second derivates of $h_T(t)$ are given by

$$h'_T(t) = -\frac{be^{a+t}}{(e^a + e^t)^2}, \tag{3.80}$$

$$h''_T(t) = \frac{be^{a+t}(e^t - e^a)}{(e^a + e^t)^3}. \tag{3.81}$$

First we examine the overall shape of $h_T(t)$, and compare it to that of $h(t)$. Note that the first derivative $h'_T(t)$ has no roots outside of the trivial case $b = 0$. The second derivative has a root at $t = a$, and like $h(t)$ there is an obvious extension to the domain $\mathbb{R}$ in which it is clear that, again like $h(t)$, the function has a sigmoid shape. This results in a very good qualitative agreement between the two functions.

Since the second derivative $h_T''(t)$ defined in Equation (3.81) has its only root at $t = a$, selecting $a < 0$ results in $h_T(t)$ being strictly convex decreasing — this is the part of the parameter space intended to model neofunctionalization. This behaviour is qualitatively equivalent to the behaviour of $h(t)$ in the case $\gamma > \gamma_{\text{crit}}$.

Similarly, when $a = 0$ the change in concavity for $h_T(t)$ occurs at $t = 0$, which is equivalent to the behaviour of $h(t)$ in the case $\gamma = \gamma_{\text{crit}}$.

When $a > 0$ there is a change in the concavity of $h_T(t)$ at $t = a$, this is the part of the parameter space intended to model subfunctionalization, and its behaviour is equivalent to that of $h(t)$ when $\gamma < \gamma_{\text{crit}}$.

Our observations in Section 3.5 show that decreasing γ below γ_{crit} increases the time t at which the point of inflection in $h(t)$ occurs. In fact, for the extended domain this behaviour generalizes such that increasing γ above γ_{crit} moves the point of inflection to the left. The same is true for decreasing a below $a = 0$ in a similarly extended domain for $h_T(t)$. Thus there is a clear correspondence between parameter a for $h_T(t)$, and γ and z for $h(t)$ (noting that γ_{crit} decreases with z), such that increasing a is roughly equivalent to increasing γ, or decreasing z.

The main advantage of a phenomenological approximation such as this is the relative simplicity of implementation in the context of biological science. The function can be computed without the construction of any matrices, and is quick and easy to implement in a scientific computing environment. As far as such an approximation goes, based on its shape properties you could likely not do much better than $h_T(t)$ for modelling subfunctionalization.

Given the overall similarity of the qualitative features of $h_T(t)$ and $h(t)$, we were interested in the possibility of inferring the more bioligically meaningful parameters of $h(t)$ based on parameters of $h_T(t)$. One approach to doing so would be to fit $h(t)$ to $h_T(t)$ by minimizing some distance measure (which we will discuss briefly) but this requires the full implementation of the underlying model for $h(t)$. As an alternative, we derive some results in support of, and describe an alternative (heuristic) procedure allowing us to choose parameters directly according to some equations which ensure that certain qualitative features of $h_T(t)$ are preserved by $h(t)$. The practical value is that it could allow a researcher to get an idea of the underlying mechanistic parameters of the more complex model while only requiring the implementation of the simpler one, which may have some utility when this forms one part of some larger scientific analysis.

First, note that $h_T(t)$ has a constant term, u which is equal to its limit as $t \to \infty$.

Thus we require

$$u = \lim_{t\to\infty} h(t) = 0, \tag{3.82}$$

in order for the limiting behaviour of $h(t)$ to match that of $h_K(t)$. If $u \neq 0$ this procedure is not likely to be of much value.

Matching the initial value of $h(t)$ to that of $h_T(t)$

The initial value of $h_T(t)$ is (assuming $u = 0$)

$$h_T(0) = \frac{be^a}{1+e^a}.$$

Since $h(0) = 2u_c$, we set $h(0) = h_T(0)$ by choosing u_c according to

$$u_c = \frac{be^a}{2(1+e^a)}. \tag{3.83}$$

Matching the location of the point of inflection of $h(t)$ to that of $h_T(t)$

Recall that the point of inflection of $h_T(t)$ occurs at the point $t = a$. Thus the remaining parameters of our model (noting that one of the two is just an integer which is, biologically speaking, unlikely to be much larger than 10) can be chosen such that

$$h_P^{z\prime\prime}(a) = 0, \tag{3.84}$$

in order to match the location of the point of inflection of the two functions.

However, finding an analytical solution for the point of inflection of $h(t)$ is not likely to be achievable, since the expression for $h''(t)$ (given in Equation (3.72)) is not very tractable. Moreover, even finding a numerical solution is difficult for large a. As per Lemma 2, we have $\lim_{t\to\infty} h''(t) = 0$. As a result, when a is large there are many parameterizations for which $h(a) \approx 0$, and these dominate the one with a point of inflection at $t = a$. This makes matching the point of inflection of the two functions unsuitable for our heuristic procedure.

Nonetheless, finding a numerical solution to Equation (3.84) is tractable for suitably small a, and this results in the point of inflection for $h_T(t)$ and $h(t)$ occurring at the same value $t = a$.

Matching the limit as $t \to -\infty$ of $h(t)$ to that of $h_T(t)$

An alternative choice for the remaining parameters which does not require the implementation of the model makes use of a novel result we derived specifically for this

purpose. As mentioned in Section 3.5, there is an obvious extension for $h(t)$ to the domain $\mathbb{R}$, and given its sigmoid-like shape we reasoned that matching the limits of the two functions as $t \to -\infty$ might restrict the behaviour of the functions sufficiently even in the usual domain to get a reasonable fit. Our numerical examination from Section 3.5 indicated that $\lim_{t\to-\infty} h(t) = 2(u_c + zu_r)$, which we prove in more general form below.

Lemma 3.

Let $X(t)$ be an absorbing CTMC with some finite state space $\mathcal{S} = \{1, ..., n\} \cup \mathcal{A}$, with $\mathcal{A}$ being the set of absorbing states, initial distribution $\underline{\alpha} = [\alpha_i]_i$, and some generator $\mathbf{Q} = [q_{ij}]_{i,j\in S}$ such that

$$\mathbf{Q} = \left[\begin{array}{c|c} \mathbf{Q}^* & \mathbf{V} \\ \hline \mathbf{O} & \mathbf{O} \end{array}\right],$$

with $\mathbf{Q}^* = [q_{ij}]_{i,j\in\{1,...,n\}}$, $\mathbf{V} = [q_{ij}]_{i\in\{1,...,n\},j\in\mathcal{A}} = [\underline{v}_j]_{j\in\mathcal{A}}$.

For any $k \in \mathcal{A}$, define

$$h_k(t) = \frac{\underline{\alpha} e^{\mathbf{Q}^* t} \underline{v}_k}{1 - \int_{u=0}^{t} \underline{\alpha} e^{\mathbf{Q}^* u} \underline{v}_k du} \text{ for all } t \in \mathbb{R}. \tag{3.85}$$

For $t \geqslant 0$, $h_k(t)$ is interpreted as the instantaneous rate of transition into state k given that the process has not yet been absorbed into state k, and if $\underline{\alpha} = \underline{e}_i$ the restriction of h_k to the domain $\mathbb{R}^+ \cup \{0\}$ is prescisely the modified-cause-specific hazard rate λ_k^i.

Then,

$$\lim_{t\to-\infty} h_k(t) = -d_m,$$

where d_m is the eigenvalue of $\mathbf{Q}^*$ with largest absolute value.

Proof.

First we consider the case where $\mathbf{Q}^*$ is diagonalizable, after which we will consider the more general case. Although the diagonalizable case is a special case of the general case which follows, the argument is structurally the same and easier to follow, so we include it for clarity.

$\mathbf{Q}^*$ diagonalizable

Since $\mathbf{Q}^*$ is diagonalizable we let

$$\mathbf{Q}^* = \mathbf{A}^{-1}\mathbf{D}\mathbf{A},$$

where $\mathbf{D}$ is a diagonal matrix of the eigenvalues of $\mathbf{Q}^*$, and denote the i^{th} eigenvector of $\mathbf{Q}^*$ by d_i.

Before making use of the diagonalization, we note that

$$\lim_{t\to-\infty} h_k(t) = \lim_{t\to-\infty} \frac{\underline{\alpha} e^{\mathbf{Q}^* t}\underline{v}_k}{1-\underline{\alpha}\left(e^{\mathbf{Q}^* t}-\mathbf{I}\right)(\mathbf{Q}^*)^{(-1)}\underline{v}_k},$$

is of indeterminent form, and so we apply l'Hôpital's rule to get

$$\lim_{t\to-\infty} h_k(t) = \lim_{t\to-\infty} \frac{\underline{\alpha} e^{\mathbf{Q}^* t}\mathbf{Q}^*\underline{v}_k}{-\underline{\alpha} e^{\mathbf{Q}^* t}\underline{v}_k}.$$

Next, making use of the diagonalization, we have

$$\begin{aligned}
\lim_{t\to-\infty} h_k(t) &= \lim_{t\to-\infty} \frac{\underline{\alpha} e^{\mathbf{A}^{-1}\mathbf{D}\mathbf{A}t}\mathbf{A}^{-1}\mathbf{D}\mathbf{A}\underline{v}_k}{-\underline{\alpha} e^{\mathbf{A}^{-1}\mathbf{D}\mathbf{A}t}\underline{v}_k} \\
&= \lim_{t\to-\infty} \frac{\underline{\alpha}\mathbf{A}^{-1} e^{\mathbf{D}t}\mathbf{A}\mathbf{A}^{-1}\mathbf{D}\mathbf{A}\underline{v}_k}{-\underline{\alpha}\mathbf{A}^{-1} e^{\mathbf{D}t}\mathbf{A}\underline{v}_k} \\
&= \lim_{t\to-\infty} \frac{\underline{\alpha}\mathbf{A}^{-1} e^{\mathbf{D}t}\mathbf{D}\mathbf{A}\underline{v}_k}{-\underline{\alpha}\mathbf{A}^{-1} e^{\mathbf{D}t}\mathbf{A}\underline{v}_k} \\
&= \lim_{t\to-\infty} \frac{\sum_j [\underline{\alpha}\mathbf{A}^{-1}]_j e^{d_j t} d_j [\mathbf{A}\underline{v}_k]_j}{\sum_l -[\underline{\alpha}\mathbf{A}^{-1}]_l e^{d_l t}[\mathbf{A}\underline{v}_k]_l}.
\end{aligned} \tag{3.86}$$

Further, denoting the dominating eigenvalue (the one with maximum absolute value) by d_{m} we divide the top and bottom of Equation (3.86) by $e^{d_m t}$ to get

$$\lim_{t\to-\infty} h_k(t) = \lim_{t\to-\infty} \frac{\sum_j [\underline{\alpha}\mathbf{A}^{-1}]_j e^{(d_j-d_{\text{m}})t} d_j [\mathbf{A}\underline{v}_k]_j}{-\sum_l [\underline{\alpha}\mathbf{A}^{-1}]_l e^{(d_l-d_{\text{m}})t}[\mathbf{A}\underline{v}_k]_l}. \tag{3.87}$$

Consider the numerator of Equation (3.87),

$$\begin{aligned}
&\lim_{t\to-\infty} \sum_j [\underline{\alpha}\mathbf{A}^{-1}]_j e^{(d_j-d_{\text{m}})t} d_j [\mathbf{A}\underline{v}_k]_j \\
&= [\underline{\alpha}\mathbf{A}^{-1}]_m d_m [\mathbf{A}\underline{v}_k]_m + \lim_{t\to-\infty} \sum_{j\neq m} [\underline{\alpha}\mathbf{A}^{-1}]_j e^{(d_j-d_{\text{m}})t} d_j [\mathbf{A}\underline{v}_k]_j \\
&= d_m [\underline{\alpha}\mathbf{A}^{-1}]_m [\mathbf{A}\underline{v}_k]_m,
\end{aligned} \tag{3.88}$$

where the final step follows from the fact that the eigenvalues of $\mathbf{Q}^*$ are necessarily negative, and that $\lim_{t\to\infty} e^{\mathbf{Q}^* t} = 0$.

Now consider the denominator of Equation (3.87),

$$\begin{aligned}
&-\lim_{t\to-\infty} \sum_l [\underline{\alpha}\mathbf{A}^{-1}]_l e^{(d_l-d_{\text{m}})t} [\mathbf{A}\underline{v}_k]_l \\
&= -\left([\underline{\alpha}\mathbf{A}^{-1}]_m [\mathbf{A}\underline{v}_k]_m + \lim_{t\to-\infty} \sum_{l\neq m} [\underline{\alpha}\mathbf{A}^{-1}]_l e^{(d_l-d_{\text{m}})t} [\mathbf{A}\underline{v}_k]_l \right) \\
&= -[\underline{\alpha}\mathbf{A}^{-1}]_m [\mathbf{A}\underline{v}_k]_m.
\end{aligned} \tag{3.89}$$

Combining equations (3.88) and (3.89) we have

$$\begin{aligned}\lim_{t\to-\infty} h_k(t) &= \frac{d_m[\underline{\alpha}\mathbf{A}^{-1}]_m[\mathbf{A}\underline{v}_k]_m}{-[\underline{\alpha}\mathbf{A}^{-1}]_m[\mathbf{A}\underline{v}_k]_m} \\ &= -d_m. \end{aligned} \tag{3.90}$$

$\mathbf{Q}^*$ not (necessarily) diagonalizable

Let

$$\mathbf{Q}^* = \mathbf{P}^{-1}\mathbf{J}\mathbf{P} \tag{3.91}$$

be the Jordan canonical form of $\mathbf{Q}^*$, with

$$\mathbf{J} = [\mathbf{J}_1 \oplus \ldots \oplus \mathbf{J}_n], \tag{3.92}$$

where n is the number of unique eigenvalues of $\mathbf{Q}^*$, and each $\mathbf{J}_i$ is a Jordan block. Here $\oplus$ represents the matrix direct sum.

Each block $\mathbf{J}_i$ is associated with a unique eigenvalue of $\mathbf{Q}^*$, denoted d_i. If d_i has algebraic multiplicity a_i, then block $\mathbf{J}_i$ has matrix size $a_i \times a_i$ and $\mathbf{J}_i$ has the form

$$\mathbf{J_i} = \begin{bmatrix} d_i & 1 & & & \\ & d_i & 1 & & \\ & & \ddots & \ddots & \\ & & & d_i & 1 \\ & & & & d_i \end{bmatrix}. \tag{3.93}$$

By the same arugment as for the diagonalizable case (applying l'Hôpital's rule), we have

$$\lim_{t\to-\infty} h_k(t) = \lim_{t\to-\infty} \frac{\underline{\alpha}\mathbf{P}^{-1}e^{\mathbf{J}t}\mathbf{J}\mathbf{P}\underline{v}_k}{-\underline{\alpha}\mathbf{P}^{-1}e^{\mathbf{J}t}\mathbf{P}\underline{v}_k}. \tag{3.94}$$

We note that,

$$\begin{aligned} e^{\mathbf{J}t} &= [e^{\mathbf{J}_1} \oplus \ldots \oplus e^{\mathbf{J}_n}] \\ &= [e^{d_1 t}\mathbf{K}_1 \oplus \ldots \oplus e^{d_n t}\mathbf{K}_n], \end{aligned} \tag{3.95}$$

where,

$$\mathbf{K}_i = \begin{bmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \frac{t^{a_i-1}}{(a_i-1)!} \\ & 1 & t & \cdots & \frac{t^{a_i-2}}{(a_i-2)!} \\ & & \ddots & \ddots & \vdots \\ & & & 1 & t \\ & & & & 1 \end{bmatrix}. \tag{3.96}$$

We can rewrite Equation (3.94) as a sum,

$$\lim_{t\to-\infty} h_k(t) = \lim_{t\to-\infty} \frac{\sum_j [\underline{\alpha}\mathbf{P}^{-1}]_j \sum_i \left[[e^{d_1 t}\mathbf{K}_1 \oplus \ldots \oplus e^{d_n t}\mathbf{K}_n]\mathbf{J}\right]_{ji} [\mathbf{P}\underline{v}_k]_j}{-\sum_l [\underline{\alpha}\mathbf{P}^{-1}]_l \sum_g [e^{d_1 t}\mathbf{K}_1 \oplus \ldots \oplus e^{d_n t}\mathbf{K}_n]_{lg} [\mathbf{P}\underline{v}_k]_l}. \tag{3.97}$$

Next, we divide the top and bottom of Equation (3.97) by $e^{d_m t}$, where d_m is the dominating eigenvalue of $\mathbf{Q}^*$. The expression on the right hand side of Equation (3.97) is then

$$\lim_{t\to-\infty} \frac{\sum_j [\underline{\alpha}\mathbf{P}^{-1}]_j \sum_i \left[[e^{(d_1-d_m)t}\mathbf{K}_1 \oplus \ldots \oplus K_m \oplus \ldots \oplus e^{(d_n-d_m)t}\mathbf{K}_n]\mathbf{J}\right]_{ji} [\mathbf{P}\underline{v}_k]_j}{-\sum_l [\underline{\alpha}\mathbf{P}^{-1}]_l \sum_g [e^{(d_1-d_m)t}\mathbf{K}_1 \oplus \ldots \oplus K_m \oplus \ldots \oplus e^{(d_n-d_m)t}\mathbf{K}_n]_{lg} [\mathbf{P}\underline{v}_k]_l}. \tag{3.98}$$

Now, we bring the limit inside the (finite) sums, and inside the matrix direct sum. Noting that e^t approachs zero much faster than any of the matrix entries as $t \to -\infty$, we have

$$\lim_{t\to-\infty} h_k(t) = \frac{\sum_j [\underline{\alpha}\mathbf{P}^{-1}]_j \sum_i [[\mathbf{0} \oplus \ldots \oplus \lim_{t\to-\infty} K_m \oplus \ldots \oplus \mathbf{0}]\mathbf{J}]_{ji} [\mathbf{P}\underline{v}_k]_j}{-\sum_l [\underline{\alpha}\mathbf{P}^{-1}]_l \sum_g [\mathbf{0} \oplus \ldots \oplus \lim_{t\to-\infty} K_m \oplus \ldots \oplus \mathbf{0}]_{lg} [\mathbf{P}\underline{v}_k]_l}. \tag{3.99}$$

If we let j_1, j_{a_m} be the first and last index associated with the block $\mathbf{K}_m$ respectively, we can write

$$\lim_{t\to-\infty} h_k(t) = \frac{\sum_{j=j_1}^{j_{a_m}} [\underline{\alpha}\mathbf{P}^{-1}]_j \sum_{i=j_1}^{j-1} [\lim_{t\to-\infty} K_m \mathbf{J}_m]_{ji} [\mathbf{P}\underline{v}_k]_j}{-\sum_{l=j_1}^{j_{a_m}} [\underline{\alpha}\mathbf{P}^{-1}]_l \sum_{g=j_1}^{l-1} [\lim_{t\to-\infty} K_m]_{lg} [\mathbf{P}\underline{v}_k]_l}. \tag{3.100}$$

If $a_m = 1$, then $\mathbf{K}_m = 1$ and $\mathbf{J}_m = d_m$, and Equation (3.100) reduces to $\lim_{t\to-\infty} h_k(t) = -d_m$, and the proof is complete. Otherwise, by carefully considering the form of K_m and J_m, we get

$$\lim_{t\to-\infty} h_k(t) = \frac{\sum_{j=j_1}^{j_{a_m}} [\underline{\alpha}\mathbf{P}^{-1}]_j \lim_{t\to-\infty}(d_m + \sum_{i=j_1}^{j-1} \frac{t^{i-1}}{(i-1)!}(1 + \frac{t d_m}{i}))[\mathbf{P}\underline{v}_k]_j}{-\sum_{l=j_1}^{j_{a_m}} [\underline{\alpha}\mathbf{P}^{-1}]_l \lim_{t\to-\infty}(1 + \sum_{g=j_1}^{l-1} \frac{t^g}{g!})[\mathbf{P}\underline{v}_k]_l}, \tag{3.101}$$

then dividing the top and bottom by t^{a_m-1} we get

$$\lim_{t\to-\infty} h_k(t) = \frac{\sum_{j=j_1}^{j_{a_m}} [\underline{\alpha}\mathbf{P}^{-1}]_j \lim_{t\to-\infty}(d_m t^{1-a_m} + \sum_{i=j_1}^{j-1} \frac{t^{i-a_m}}{(i-1)!}(1 + \frac{t d_m}{i}))[\mathbf{P}\underline{v}_k]_j}{-\sum_{l=j_1}^{j_{a_m}} [\underline{\alpha}\mathbf{P}^{-1}]_l \lim_{t\to-\infty}(t^{1-a_m} + \sum_{g=j_1}^{l-1} \frac{t^{g-a_m}}{g!})[\mathbf{P}\underline{v}_k]_l}. \tag{3.102}$$

After taking the limit we are left with

$$\begin{aligned}\lim_{t\to-\infty} h_k(t) &= \frac{[\underline{\alpha}\mathbf{P}^{-1}]_{a_m} \frac{d_m}{(a_m-1)!} [\mathbf{P}\underline{v}_k]_{a_m}}{-[\underline{\alpha}\mathbf{P}^{-1}]_{a_m} \frac{1}{(a_m-1)!} [\mathbf{P}\underline{v}_k]_{a_m}} \\ &= -d_m. \end{aligned} \tag{3.103}$$

□

Remark 7.

Crossman [28] showed that the hazard rate associated with an absorbing birth-death process is bounded below by the negative of the dominating eigenvalue of $\mathbf{Q}^*$*. This provides an interesting connection between the modified-cause-specific hazard rate and the hazard rate, at least when the process meets the requirements of the result in [28].*

Applying Lemma 3 to our pseudogenization rate, we have

$$\lim_{t\to-\infty} h(t) = 2(u_c + zu_r).$$

$h_T(t)$ also has an obvious extension to the domain $\mathbb{R}$, though evaluating its limit is somewhat more straightforward. Considering $h_T(t)$ extended to domain $\mathbb{R}$ we have,

$$\lim_{t\to-\infty} h_T(t) = u + b,$$

Thus (assuming $u = 0$) we can choose u_r and z such that

$$b = 2(u_c + zu_r). \tag{3.104}$$

Assuming u_c was chosen according to Equation (3.83), we can substitute this into Equation (3.104) to get

$$b = \frac{be^a}{(1+e^a)} + 2zu_r,$$
$$\text{i.e. } u_r = \frac{b}{2z}\left(1 - \frac{e^a}{(1+e^a)}\right). \tag{3.105}$$

Since z is an integer, Equation (3.105) has only finitely many solutions for fixed a, b, and only a handful of them correspond to biologically realistic parameters (although we will see that choosing large z gives the best fit).

Although the function's value for negative t has no physical relevence, taken together with the value at $t = 0$ it provides a reasonable amount of restriction on the of the behaviour of $h(t)$ over all t. Notably, choosing u_r according to Equation (3.105) and u_c according to Equation (3.83) results in u_c and u_r increasing and decreasing with a respectively, so that their ratio γ is decreasing in a. This is consistent with the observation that increasing a results in qualitative behaviour similar to decreasing γ. Further, u_r is decreasing in z, and hence γ is decreasing in γ_{crit}, which is also consistent with this observation. So, at least in terms of reproducing the qualitative behaviour of $h_T(t)$, selecting u_r and u_c according to equations (3.105) and (3.83) is an appropriate choice.

Thus, the proposed heuristic procedure for finding parameters of $h(t)$ to match the behaviour of $h_K(t)$ for some known parameters is to choose u_c according to Equation (3.83), and find a range for u_r by solving Equation (3.105) for several values of

z. This provides a rough estimate of the mutation rates associated with a particular parameterization of $h_K(t)$ according to the model described in this chapter.

To test the quantitative performance of the procedure we computed $h_T(t)$ on a grid of parameters with a in the region $[-10, 20]$ and b in the region $[0.1, 10]$ with intervals of 0.1 for both, together with $h(t)$ for each of $z = 2, 3, 4, 5$ (with the other parameters chosen as described). The two functions were evaluated at 100 equally spaced points on in the region $0 < t < t_{\max} = \max\{|2a|, 10\}$. This choice of interval for t ensured that most of the behaviour of $h_T(t)$ was captured (in the sense that it was near to 0 by the end of the interval).

Remark 8.

In retrospect a better choice would have been an interval such that $h_T(0)/h_T(t_{\max})$ was constant, which would have yielded a fairer comparison between different parameters. However, given the exploratory nature of this investigation the chosen interval was sufficient for our purposes, and we did not feel it necessary to re-run the analysis.

For fixed z, t we calculated $|h_T(t) - h(t)|/h_T(t)$ as a measure of the relative distance between $h_T(t)$ and $h(t)$, which we averaged over the 100 time points. This gives a measure of the average relative difference between $h_T(t)$ and $h(t)$, which we then averaged over z. The results of this examination are shown as a heatmap in Figure 3.10.

For a very narrow range around $b = 2$ with $a \geqslant 1$ the distance measure was on the order of 0.01, which corresponded to a very close fit — Figure 3.11 shows an example from this region. The rest of the parameter space was associated with fairly poor fits, with those for which $b << 2$ the worst. Figure 3.13 shows one of the least successful examples (as measured by the average relative distance on the chosen interval), while Figure 3.14 shows a fairly typical one. Further investigation revealed that the fit became very close as z became very large for $b = 2$ for any a. We tested $z = 50, 100, 150, 200$ for $a = -10, -9, ..., 10$, as well as 10 randomly chosen a from the same interval and $b = 1.9, 2, 2.1$. We found that increasing z always yielded a better fit (on the order of average relative distance 0.001 for $z = 200$), and that the fit for $b = 2$ was generally better than for the other two values. It appears as though somewhere near to $b = 2$ $h_T(t)$ may be an exact solution for $h(t)$ in the limit as $z \to \infty$ with u_c and u_r chosen according to Equations (3.83) and (3.104) respectively, though we would not give this too much credence based on this analysis. Whether or not the exact solution hypothesis is true, it is certainly the case that $h(t)$ with large z can be fit very closely to $h_T(t)$ for $b = 2$ by our heuristic procedure.

The final part of our analysis was quantitative, testing the potential to fit $h(t)$ to

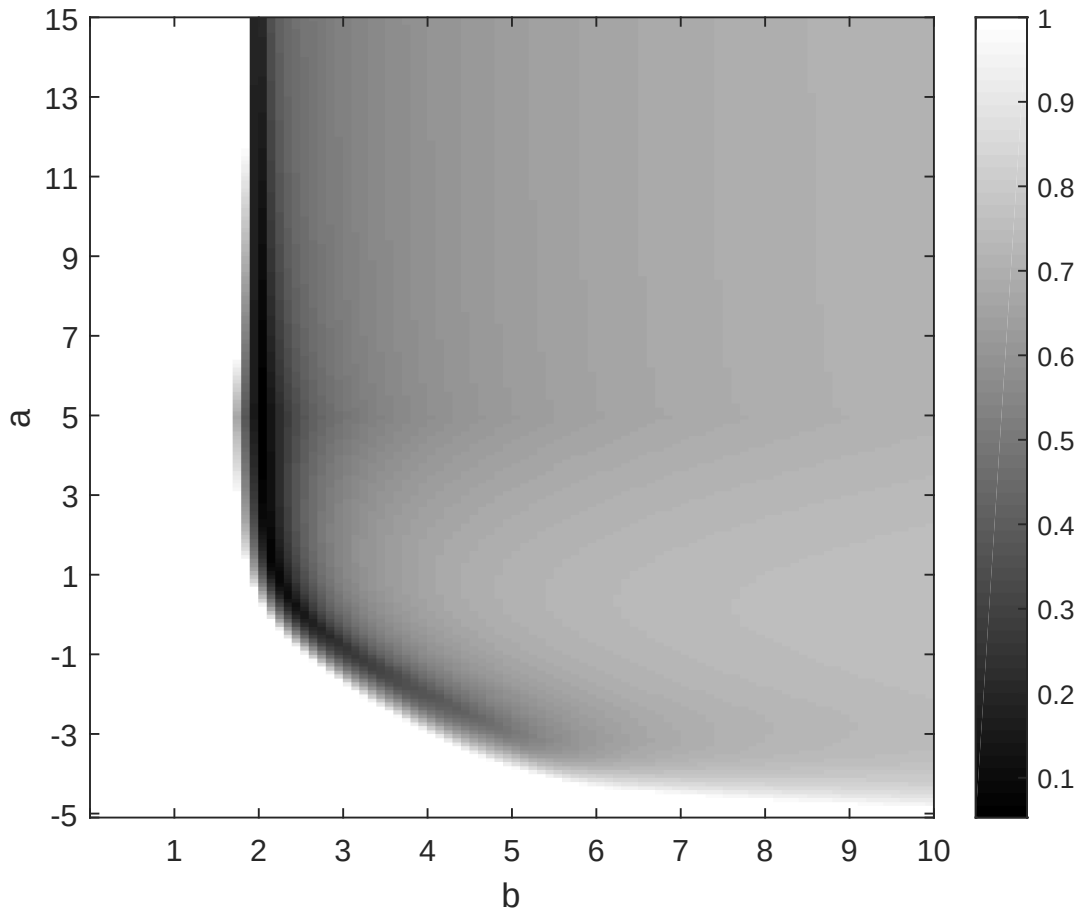

Figure 3.10: A heatmap showing the average relative difference between $h_T(t)$ (for various values of a, b) and $h(t)$ averaged over $z = 1, ..., 5$ for the parameters inferred by our heuristic. We imposed a maximum of 1 on the colourmap, since a difference of 1 or more represented substantial divergence between the two functions.

$h_T(t)$ numerically (using the same distance measure as above) for a grid of paramter values with a in the region $[-5, 5]$ and b in the region $[0.1, 10]$, again with intervals of 0.1 for both. For each of $z = 2, 10, 20$ we used the MATLAB function *fminsearch* (which uses a Nelder-Mead simplex) to find parameters minimising the average relative distance between $h(t)$ and $h_T(t)$ at 100 equally spaced points in the region $0 < t < \max\{|2a|, 10\}$. The search was well behaved, meeting the stopping criterea in all but a handful of instances where the maxmium allowed iterations was exceeded, together with consistent results across most of the grid, this suggests that the search was likely choosing the true optimum parameters to minimize the distance measure on the chosen interval.

The closest fit overall was for $b = 2, a = 1.8, z = 20$, and the fit was good for most

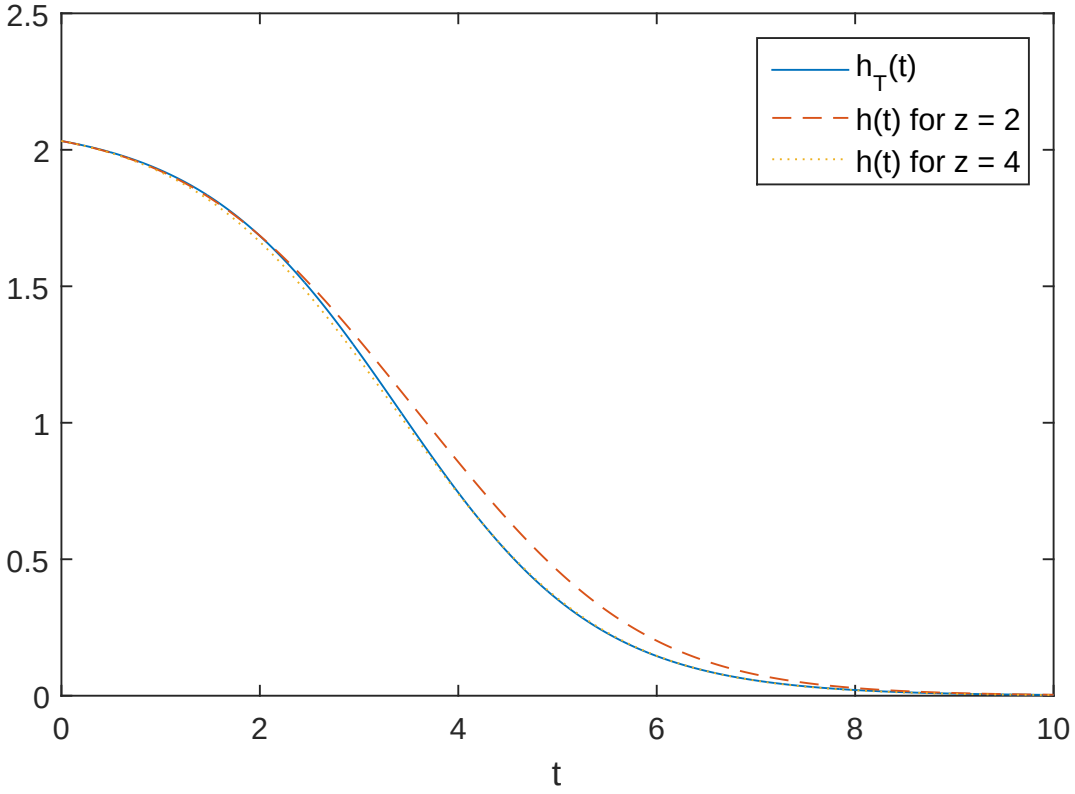

Figure 3.11: $h_T(t)$ for $a = 3.4, b = 2.1$ (average distance measure 0.07) together with the fitted $h(t)$ for $z = 2$ and $z = 4$. At this scale, the plot of $h(t)$ for $z = 4$ falls roughly on top of that of $h_T(t)$.

a in the region around $b = 2$, with very similar parameter estimates to the heuristic procedure in the region $a > 1$. The fit was also good for any $b > 1$ with a in the region from around -1 to 2, which clearly demonstrates that $a < 0$ cannot be uniquely associated to neofunctionalization, but outside of these regions the fit was not very close. The best fit was achieved for $z = 20$ in 70% of estimates, and 77% of estimates for which the average relative distances was less than 0.1. It is possible that better fits could be achieved by considering more values of z, in particular, it appears that $h(t)$ fits more closely to $h_T(t)$ as z becomes larger for most parameters, but we did not investigate this further. Figure 3.12 shows a heatmap of the minimum (over z) average relative distance of $h(t)$ from $h_T(t)$ achieved by the optimization.

Overall $h_T(t)$ is qualitatively and, to a lesser extent, quantitatively similar to our pseudogenization rate function $h(t)$. Given that $h_T(t)$ is essentially a purely qualitatively inspired phenomenological approximation to the hazard rate for pseudogenization under subfunctionalization, the agreement between it and our exact function derived from a model of the underlying mechanics is fairly impressive. However, the fit was

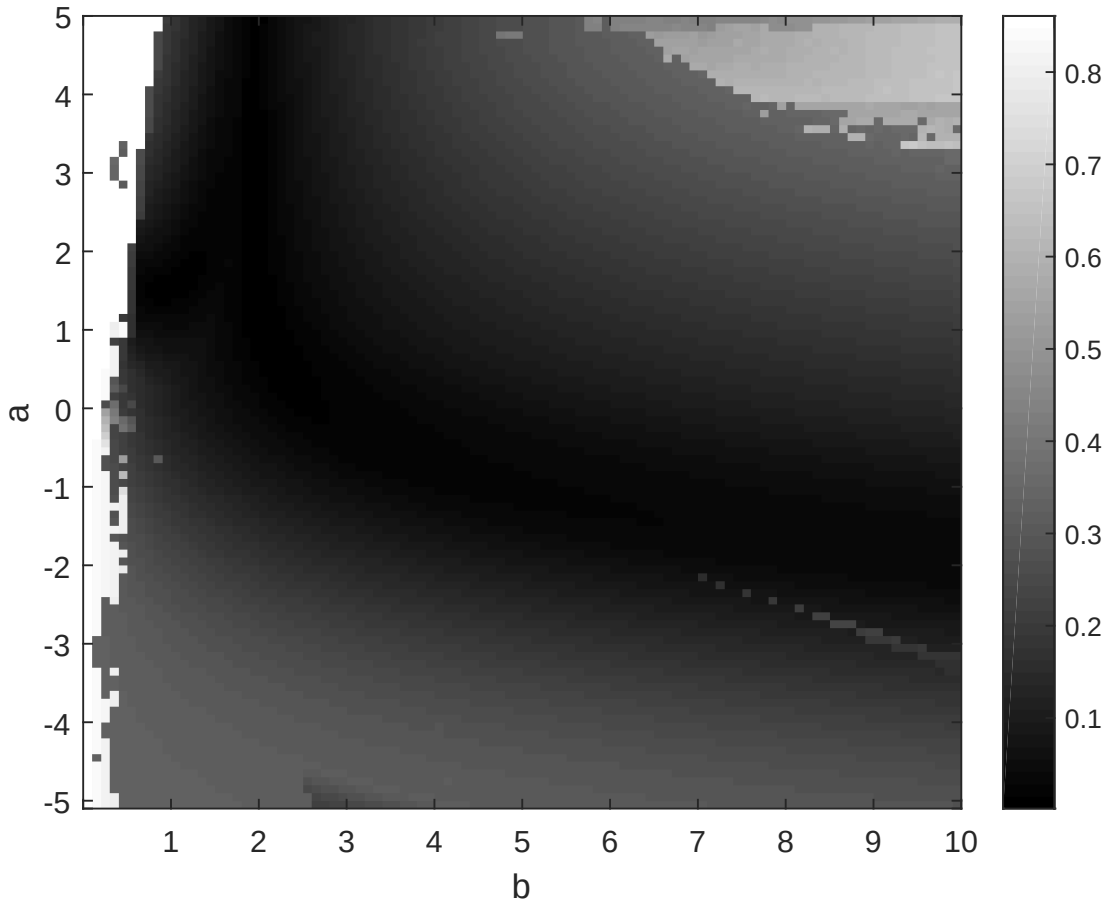

Figure 3.12: A heatmap showing the average relative difference between $h_T(t)$ (for various values of a, b) and $h(t)$ averaged over $z = 1, ..., 5$ for the parameters inferred by our heuristic. We imposed a maximum of 1 on the colourmap, since a difference of 1 or more represented substantial divergence between the two functions.

also very close for the region of the parameter space thought to correspond not to sub- but to neofunctionalization, which highlights the need for rigorous mathematical analysis. Further, $h(t)$ seems to fit best to $h_T(t)$ as $z \to \infty$, which is far from being biologically realistic.

Our attempt to find a similarly qualitatively inspired procedure to fit $h(t)$ to $h_T(t)$ was not very successful; notably the procedure provided very close fits around $b = 2$, which from our numerical work appears to in fact be the place where the functions agree most closely. On the one hand, this result seems quite novel, but on the other hand if the procedure was going to work anywhere, it was likely to be for those parameters for which the two functions agreed most closely. We had initially intended to match the initial value, point of inflection (in the extended domain $\mathbb{R}$), and limit as $t \to \infty$ of the

functions, but could not find an analytical solution for when the points of inflection agreed. Instead, we matched the limit as $t \to -\infty$ of the two functions extended to the domain $\mathbb{R}$. The rationale was that, given the functions' sigmoid shape, matching the values of this limit (together with initial value, and limit as $t \to \infty$) might restrict the behaviour even in $\mathbb{R}^+ \cup \{0\}$ sufficiently to achieve a reasonable fit. Ultimately for most a and b attempting to infer u_r and u_c using the heuristic procedure was fruitless. Numerical estimation to associate parameters of $h_T(t)$ to those of $h(t)$ may have some merit, but the fit we achieved was not sufficiently close for most parameters to justify proceeding, especially given the preference for unrealistically large z.

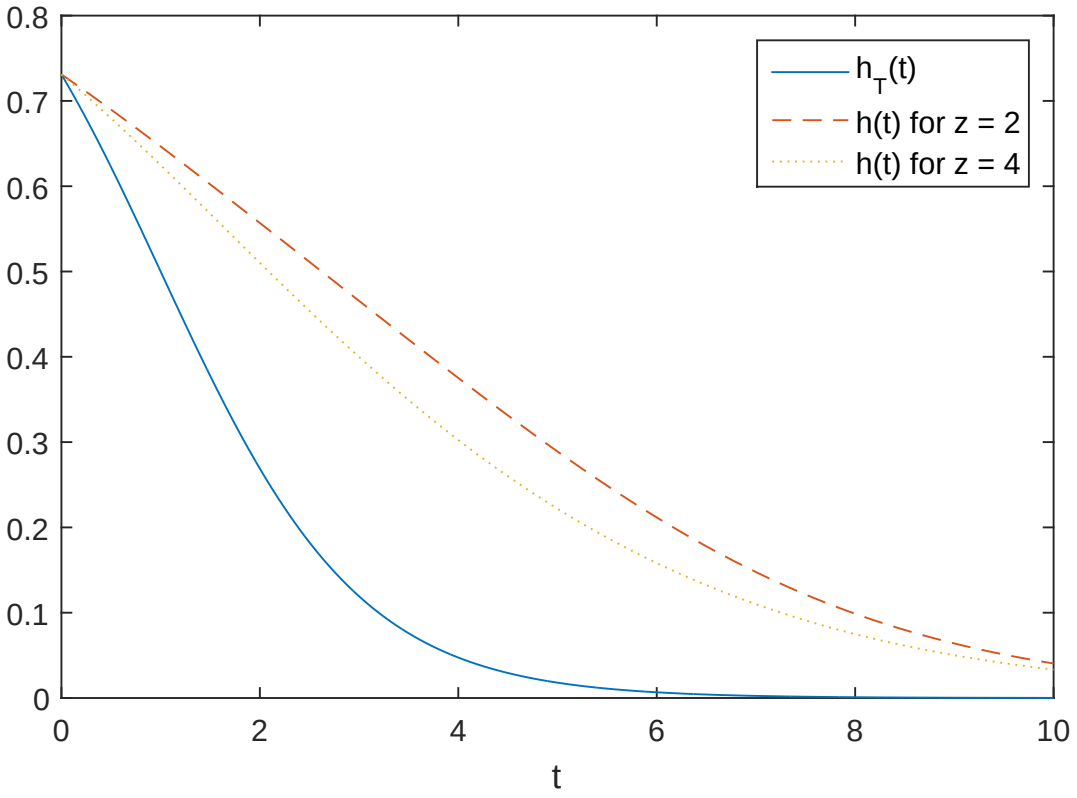

Figure 3.13: $h_T(t)$ for $a = b = 1$ (average distance measure 44.62) together with the fitted $h(t)$ for $z = 2$ and $z = 4$. $h(t)$ approaches its limit much more slowly than $h_T(t)$ on the measured interval. There are three orders of magnitude difference in the value of $h_T(t)$ and $h(t)$ by the time $t = 10$, which was heavily penalized by the relative difference measure.

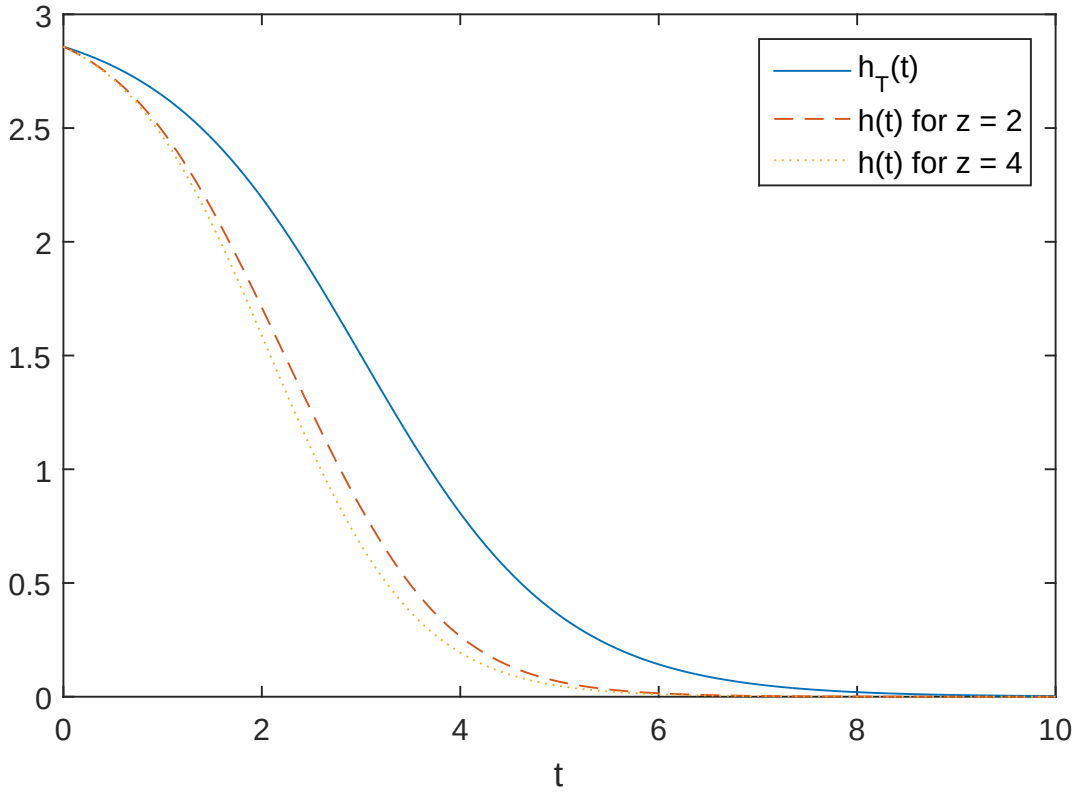

Figure 3.14: $h_T(t)$ for $a = b = 3$ (average distance measure 0.68) together with the fitted $h(t)$ for $z = 2$ and $z = 4$. Although the fit was not very good, the functions did approached their limits at similar rates on the measured interval so that the relative distance (compared to the example with $a = b = 1$) was fairly consistent.

3.7 Extending the model to a population of duplicate pairs via a Poisson birth process

In [109] we fit the model introduced in this chapter to a particular dataset handled by Hughes and Liberles [54]. The data essentially amounts to counts of the number of surviving duplicates falling into various age brackets (we give more details in Section 3.8), but no information is available to determine whether a particular duplicate pair has undergone subfunctionalization. The number of duplicates surviving to time t depends on the survival process, which we model via Equation (3.35), as well as on the duplication process. To this end, we assume that the duplication process is Poisson, and derive the likelihood of the data D given the Poisson duplication process and the survival process implied by our model.

The age of the duplicates was determined by Hughes and Liberles by proxy via the

expected number of silent substitutions per silent site. That is, the expected number of mutations that will occur without effecting the expression of the gene, scaled by the reciprocal of number of sites at which such mutations can occur. An estimate of the (scaled) number of such mutations was made by Hughes and Liberles [53] for each duplicate pair, which were then put into bins of width $0.01s$, where s is the expected number of silent substitutions per silent site — this can be interpreted as the expected time for one out of one hundred silent sites to undergo silent substitution.

Let N be a Poisson random variable with parameter β_0 which counts the number of duplication events occurring in a time interval of length $0.01s$ (chosen to match the bin size of the data). Also, let $Y(t)$ be a random variable tracking the number of duplicates which have survived to the current time, having been duplicated a time t in the past. It follows that

$$\begin{aligned}
P(Y(t)=y) &= \sum_{n\geqslant y} P(Y(t)=y \mid N=n)P(N=n) \\
&= \sum_{n\geqslant y} (1-\tilde{F}(t))^y \tilde{F}(t)^{n-y} \binom{n}{y} P(N=n) \\
&= (1-\tilde{F}(t))^y \sum_{n\geqslant y} \tilde{F}(t)^{n-y} \left(\frac{n!}{y!(n-y)!}\right) \frac{\beta_0^n}{n!} e^{-\beta_0} \\
&= \frac{(1-\tilde{F}(t))^y e^{-\beta_0}}{y!} \sum_{n\geqslant y} \frac{\tilde{F}(t)^{n-y}\beta_0^n}{(n-y)!} \\
&= \frac{(1-\tilde{F}(t))^y e^{-\beta_0}}{y!} \beta_0^y \sum_{n\geqslant 0} \frac{(\tilde{F}(t)\beta_0)^n}{n!} \\
&= \frac{(1-\tilde{F}(t))^y e^{-\beta_0}}{y!} \beta_0^y e^{\tilde{F}(t)\beta_0} \\
&= \frac{((1-\tilde{F}(t))\beta_0)^y}{y!} e^{-\beta_0(1-\tilde{F}(t))}, \qquad (3.106)
\end{aligned}$$

Note that (3.106) defines a nonhomogeneous Poisson random variable with parameter

$$\beta(t) = \beta_0(1-\tilde{F}(t)). \qquad (3.107)$$

We can then use a maximum likelihood method to estimate the parameters u_r, u_c and z of the subfunctionalization model together with the parameter β_0 of the duplication process. The log likelihood of parameter set $\underline{\theta} = [u_r, u_c, z, \beta_0]$ given data D is given by,

$$log(L_\theta \mid D) = \sum_i D_i \log(\beta(s_i)) - \beta(s_i) - \Gamma \log(D_i+1), \qquad (3.108)$$

where D_i is the count in the i^{th} bin of the data set, and s_i is the associated cumulative

number of silent substitutions per silent site (which, to reiterate, is a proxy for the age of the duplicates).

3.8 Fitting the model to genome data

In this section we fit the model to the data set analyzed in Hughes and Liberles [53]. The data consists of counts of the number of duplicate pairs in several genomes with corresponding estimates of the cumulative number of silent substitutions per silent site, binned in intervals of length $0.01s$, where s is the cumulative number of silent substitutions per silent site. The silent substitutions can be used as a proxy for time, and the intervals of length $0.01s$ represent on average 1.1 million years [53]. Hughes and Liberles [53] tested the quality of the alignments by comparing the mean and median fraction of alignment columns which were gap free. They concluded that the alignments for the four species *M. musculus, R. norvegicus, H. sapiens and C. familiaris* were of high quality, and these are the data sets we will examine here. In [53] Hughes and Liberles fit a Weibull function to the survival data. From the parameters of the fitted survival function, they inferred that the hazard rate must be convexly decreasing, concluding that the data was more consistent with neo- than subfunctionalization. Here we assume that the underlying duplication process is a Poisson process, and fit the survival function derived from our model of subfunctionalization to the data directly as discussed above in Section 3.7.

We computed maximum likelihood estimates $\hat{\theta}_z = [\hat{u}_r, \hat{u}_c, \hat{\beta}_0]$ for each z from 2 to 20 for four mammalian genomes. We call the best of these z's (in terms of likelihood) $\hat{z}$, with the understanding that this is not a true maximum likelihood estimate, since we restricted $\hat{z} \in \{2, 3, ..., 20\}$. We chose this truncation because it is unlikely that a gene would have in excess of 20 regulatory regions [53]. The case $z = 1$ is excluded, as subfunctionalization cannot occur in this case, and the survival model reduces to an exponential survival function with parameter $2(u_c + u_r)$ when $z = 1$ or $u_r = 0$. The possibility of this exponential survival function is already accounted for, for each z, by the fact that u_r is free to be chosen equal to 0.

The ratio $\gamma = u_r/u_c$ and z appear to be strongly correlated in the MLEs, as shown in Figure 3.15. A power law relation between γ and z appeared to fit quite well, with $R^2 > 0.97$ for each of the four genomes.

We compared the fit of our survival function (3.35) against Weibull and exponential functions using relative likelihood via the AIC (given in Equation (2.70) of Chapter 2). For all four genomes, our model outperformed the exponential function, but was itself

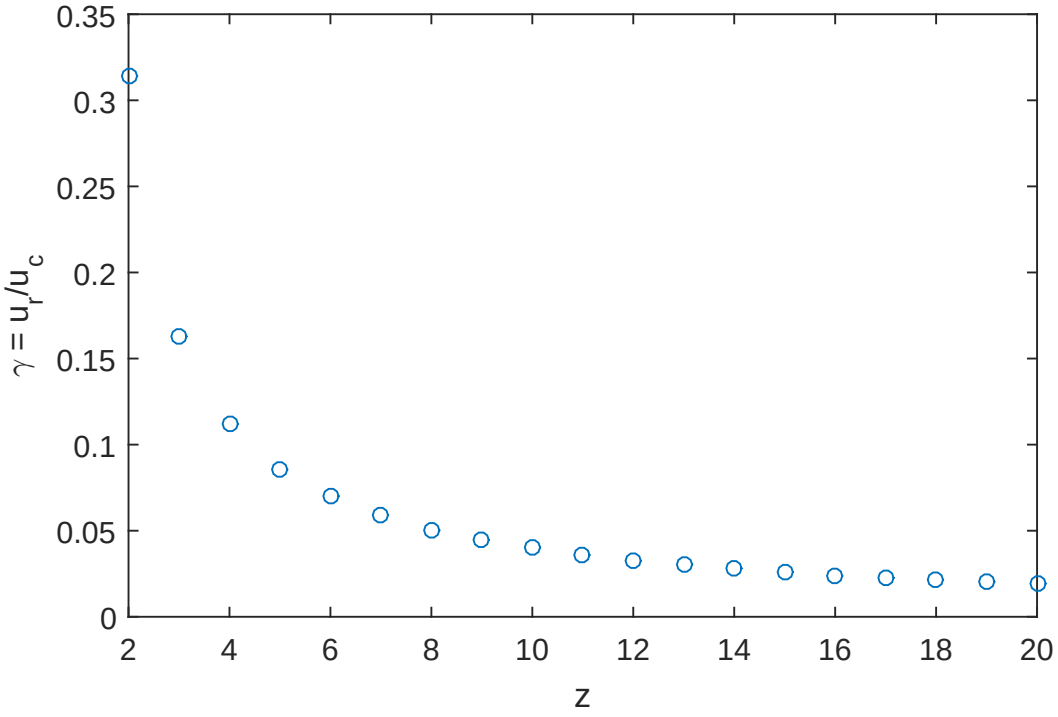

Figure 3.15: Maximum likelihood estimates for $\gamma = u_r/u_c$ for each $z = 2, 3, ..., 20$ for *Mus musculus*.

outperformed by the Weibull function in the rat, mouse and human genomes. In the canine genome there was insufficient evidence to choose between the Weibull function and the survival function derived from the model.

Mechanistic models can contain parameters that are part of the generative process but add little to data fitting, sometimes resulting in less support for mechanistic models when compared to simpler models, even when the mechanistic models give more accurate inference of the underlying process as judged by the accuracy of parametrization (see Liberles et al. [79]). With this in mind, we move forward with analysis of the results of fitting the mechanistic subfunctionalization model to genomic data. The analysis of mechanistic parameters is conditional on the generative process being what is modeled, namely subfunctionalization.

To estimate the relative rates of mutation $\gamma = u_r/u_c$ together with the mean number of duplicates per $0.01s$, β_0, we fixed $z = \hat{z}$ and computed e^2 likelihood intervals for each of the parameters using the profile likelihood approach. We also calculated e^2 likelihood intervals for z using the values of the MLE. The e^2 likelihood interval is defined as $\{\underline{\theta} : L(\underline{\theta}|D) \geqslant e^{-2}L(\hat{\underline{\theta}}|D)\}$. In the regular asymptotic case e^2 likelihood intervals are equivalent to 95.4% confidence intervals [52]. Since the shape of the profile likelihood

	Rattus norvegicus			*Mus musculus*		
	Lower	MLE	Upper	Lower	MLE	Upper
u_c	2.24	3.04	3.68	18.03	20.07	22.33
u_r	0.00	0.67	2.41	2.87	3.26	3.69
β_0	186.63	204.04	221.06	633.00	680.84	731.62
z	2	2	20	3	3	5
	Homo sapiens			*Canis familiaris*		
	Lower	MLE	Upper	Lower	MLE	Upper
u_c	12.44	14.71	17.43	6.36	7.74	9.25
u_r	2.52	3.11	3.80	1.30	2.39	3.45
β_0	315.55	348.11	384.05	114.12	129.07	145.77
z	3	3	5	2	2	20

Table 3.1: Maximum likelihood estimates and e^2 likelihood intervals for four species. The e^2 likelihood intervals can be regarded as approximate 95% confidence intervals (although the reader will note that they are not necessarily symmetric).

is approximately normal (for example, see Figure 3.16), it is reasonable to regard these intervals as approximate 95% confidence intervals. The results are summarised in Table 3.1.

There were some identifiability issues in fitting the model to this data, in particular, for the *Rattus norvegicus* and *Canis familiaris* genomes, we were able to find good relative likelihood scores for any $z = 2, 3, ...20$, which prevents us from reliably estimating z for these two genomes. Together with the close correlation between z and $\gamma = u_r/u_c$ this could be overcome by fixing one or more of the parameters using some outside analysis. In both cases, the maximum likelihood estimate for z was $\hat{z} = 2$.

To test identifiability, we ran some simulations using parameters similar to those estimated for the rat genome. We simulated bins of data identical to those in the data, i.e. 30 bins corresponding to $0.01s, 0.02s, ...0.3s$, and found that the parameters of the model were difficult to recover in this case, with the MLE value for z varying between runs with the same parameters. In some runs, even when z was fixed to the correct value used in the simulation u_r and u_c were not able to be accurately recovered, with $\gamma = u_r/u_c$ varying from around 0.05 to 0.3 (true value 0.2) in the handful of simulations we ran. The true parameters fell within the e^2 likelihood intervals, but it was not until we increased the number of intervals to 100 that we started to get

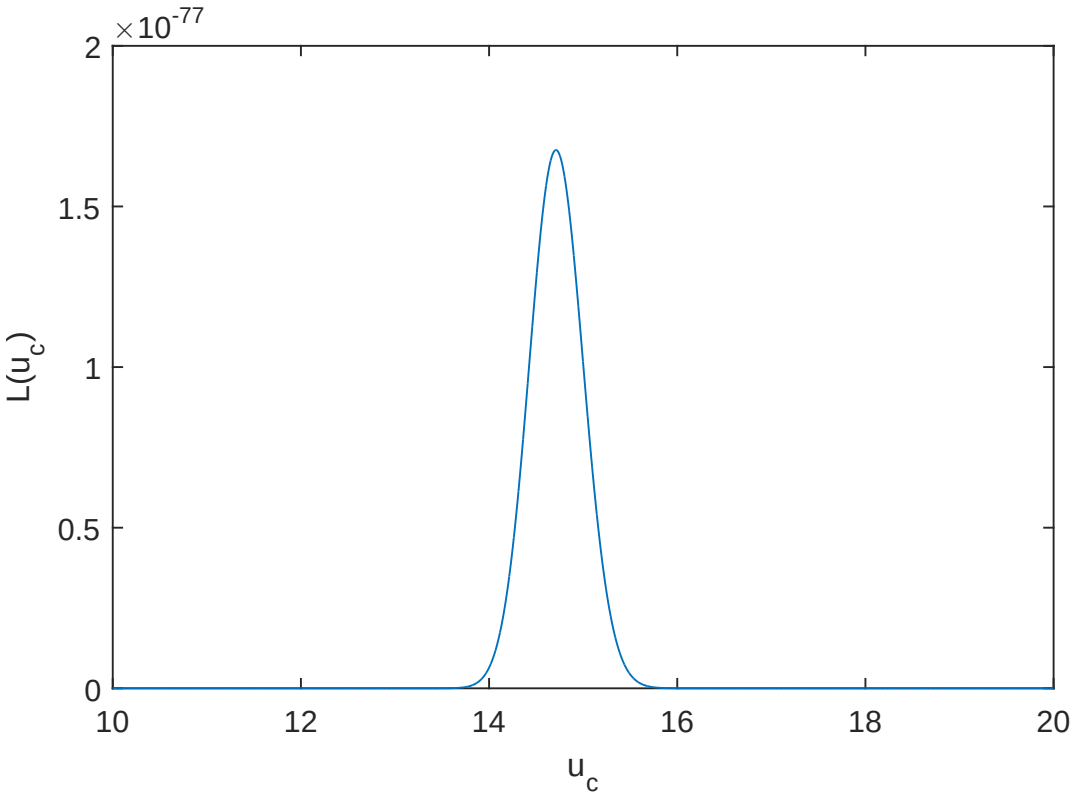

Figure 3.16: Profile likelihood $L(u_c)$ for u_c in the *Homo sapiens* genome.

reliable recovery of the simulation parameters.

The simulation analysis provided some insight into the relatively unstable results for the rat and dog genomes. We suspect that the combination of low count values (and hence low β_0), together with the relatively low estimated mutation rate, leads to an overall lack of information in the data for these two genomes compared to the others, and hence the difficulty pinpointing parameters. Based on the results of the simulations we ran, we expect that the likelihood intervals for these genomes are somewhat reliable, while the maximum likelihood estimates themselves are probably not very precise.

We also note that for the rat genome the value $u_r = 0$ was within the established likelihood intervals. In this case, the survival function for our model reduces to an exponential survival function with parameter $2u_c$, which can not really be seen as a model for subfunctionalization at all, since subfunctionalization will never occur, and the pseudogenization rate is flat.

For the *Homo sapiens* and *Mus musculus* genomes the maximum likelihood estimate

for z was $\hat{z} = 3$ in both cases, with $z = 3, 4, 5$ falling in the e^2 likelihood interval. In these two cases the higher mutation rate estimates, together with larger counts of duplicates are suggestive of more informative data, and the results are in-line with this suggestion. We expect the maximum likelihood parameter estimates to be more reliable in these cases.

Note that we model the evolution of a *pair of gene duplicates*, and thus our estimates implicitly assume that all of the duplicates in the genomes analyzed have the same parameters as each other. That is, the maximum likelihood estimate $\hat{z}$ is an estimate for the number of regulatory regions each gene has assuming they all have the same number of regulatory regions. Similarly, the estimates for u_c and u_r assume a consistent rate of mutation throughout all of the genes in the data set.

These assumptions are inherent to the application of models at the level of individuals (or in this case, pairs of individuals) to larger data sets, however the importance of these assumptions is particularly apparent when considering structural parameters such as z. In the absence of parameter z, we could think of the Poisson rates u_c and u_r as measuring an average mutation rate across the genome, however, since z is a non-stochastic, structural parameter of the model there is no similar interpretation for the number of regulatory regions. With this in mind, we can think of u_c and u_r as average mutation rates given that all duplicates examined have exactly z regulatory regions.

In order to relax this assumption, we computed analogous maximum likelihood estimates for randomly distributed Z using a truncated ($2 \leqslant Z \leqslant 20$) Poisson distribution with parameter α, given by

$$P(Z = z) = \frac{\alpha^z}{z!} e^{-\alpha} \left(\sum_{k=2}^{20} \frac{\alpha^k}{k!} e^{-\alpha} \right)^{-1}, \tag{3.109}$$

which resulted in distributions where the majority of the weight was at the lower end of the truncation, $Z = 2$. However, this result should be viewed with care, as the procedure is biased in favour of results which place the majority of the weight around the points of truncation, $Z = 2$ and $Z = 20$. This is because having the majority of the weight on a single value of Z allows for the parameters β_0, u_r, u_c to be chosen so as to best fit the particular value of Z, giving a distinct advantage over more evenly weighted distributions.

3.9 Discussion

Our analysis of the four mammalian genomes (using data handled by Hughes and Liberles [53]) suggests that (subject to our modeling assumptions) gene duplicates most likely have only a few regulatory regions, and that the rate of mutation in these regulatory regions is on the order of 10 times smaller than the rate of mutation in the coding region, which is suggestive of the relative mutational target sizes. This is the first model-based estimate of the number of regulatory regions in gene duplicates. The estimates, based upon an assumption of duplicate gene preservation through the subfunctionalization process, are in-line with the conventional thinking. Force et al. [42] suggested that the ratio of coding to regulatory mutations should be about 0.1 to 0.7, and Hughes and Liberles [53] suggested that between 2 and 12 regulatory regions were realistic. Mechanistic characterization of mutational potentials in protein-coding genes from molecular-level analysis can add additional insight into these parameterizations.

All of the parameters within our e^2 likelihood intervals (with the exception of $u_r = 0$ for the *Rattus norvegicus* genome) were consistent with a convex declining hazard rate, but we would be weary of inferring that this implies the dominant mode of subfunctionalization is such. Besides the assumptions inherent in the model, the analysis of empirical data relies also upon the assumption that the mode of preservation for all genes in the genome is subfunctionalization of the regulatory regions. The effect of duplicate pairs becoming fixed by processes other than regulatory subfunctionalization, such as neofunctionalization, or subfunctionalization of the coding sequences, would be absorbed in the parameters of our model, potentially biasing the estimates.

Subfunctionalization of the coding sequences could be well handled in some instances by the same model with slightly different interpretation (where the regulatory regions are thought of as blocks of the coding region which can be nonfunctionalized without the total loss of function). For example, regulation mediated by post-translational modification of specific amino acids might occur with regulatory region-like dynamics. On the other hand, subfunctionalization of coding sequences is less likely to be a neutral process — for example, subfunctionalization from a ligand-binding generalist to a pair of specialists that are specific to an individual or set of ligands might require selection to attain that specificity (see Liberles et al. [80] for further discussion of this point). The presence of genes preserved by neofunctionalization is likely (if anything) to bias our estimates towards lower values of γ, reflective of the convex hazard rate associated with neofunctionalization. However, in our analysis of neofunctionalization, discussed in Section 4.1, we find that even very high rates of neofunctionalization probably attribute very little such bias.

Our analysis partially contradicts Hughes and Liberles [53] characterization of subfunctionalization by an initially constant, and then broadly decreasing, concave hazard function which has been adapted by subsequent works [66; 117; 54]. The intuition behind this characterization can be explained by thinking of the initial period of constant hazard rate as corresponding to the waiting time for the first mutation. After this first mutation the unaffected gene will be selectively protected against pseudogenization, and hence there is a sharp decline in the hazard rate (from $2u_c$ to u_c in terms of both the model discussed in [53] and the model discussed in this paper). Once the first mutation has fixed, there is now an opportunity for an additional mutation to lead to subfunctionalization, in which case the rate of pseudogenization will decrease from u_c to 0. The probability that subfunctionalization occurs before time t is rapidly increasing with t, and this leads to the concave decline in the hazard function.

In contrast, the pseudogenization rate $h(t)$ (Equation (3.34)), for the parameter sets of primary biological interest, is best characterised by a sigmoid shape, or by an exponential-like shape. $h(t)$ may include a period of concave decline, but it always includes a period of convex decline, and for many parameter sets there is no concave decline at all.

For $\gamma < \gamma^z_{\text{crit}}$ the change in concavity occurs for some $t > 0$, and we see a short or long period of concave decline followed by convex decline (see Figure 3.5a). This essentially agrees with the characterization of Hughes and Liberles [53]. In fact, Fig. 7 of [53] shows a period of convex decline in the mean hazard rate when averaging over certain distributions of the number of regulatory regions z. However, the present work shows that even with a fixed number of regulatory regions z, a change in concavity will occur in the hazard rate $h(t)$. This suggests that the period of convex decline is more fundamental to subfunctionalization than suggested by the characterization of [53], which focused particularly on the period of concave decline.

For $\gamma > \gamma^z_{\text{crit}}$ the difference in our hazard rate and the characterization of Hughes and Liberles [53] is more stark, and warrants a careful reconsideration of the biological intuition. In this case, the hazard rate $h(t)$ is convexly decreasing for all $t > 0$, much like an exponential decay (see Figure 3.5b), and this is the shape which is supported by our analysis of the four mammalian genomes. Whens $\gamma > 1$, a period of increase in the pseudogenization rate is possible, but the associated relative rates of mutation are not generally considered to be realistic [42].

The prediction of a convex decay comes from the same mechanics which motivate, and (for certain parameter sets) give rise to the concave characterization. Thus, we suggest some new intuition for duplicates that have a large nonfunctionalizing mutation rate

in the regulatory regions.

Thinking of $h(t)$ as a weighted average as per Proposition 20 in Section 3.5 (restated below for convenience), we consider the evolution of a duplicate pair with moderately large nonfunctionalizing mutation rate in the regulatory regions — on the same order per regulatory region as for the entire coding region, but not significantly greater than it. Two important features are then apparent which explain the convex decline of the hazard rate $h(t)$. Recalling Equation (3.47) from Section 3.5, we have

$$h(t) = 2u_c p_{\{0\}}(t) + u_c p_{\{1,\ldots,z-2\}}(t) + (u_c + u_r) p_{\{z-1\}}(t) + 0 p_{\{S\}}(t). \tag{3.47}$$

First, there is a high probability that an initial nonfunctionalizing mutation in the regulatory region occurs in a short time. Correspondingly, the u_c term increases rapidly, balanced by a decrease in the $2u_c$ term. This results in an initial, rapid decrease in the rate of pseudogenization. Second, once this first mutation has fixed, there is a very high rate of subfunctionalization (rate of transition to S). Thus, the 0 term increases rapidly, balanced by decrease in the u_c and $u_c + u_r$ terms. Any non-pseudogenizing mutation after the first and before the $z - 1^{\text{th}}$ is equally likely to lead to subfunctionalization as it is to eliminate another region of the already-mutated copy. As such, the $(u_c + u_r)$ term increases (balanced by decrease in the u_c term) no faster than the 0 term, and the pseudogenization rate continues to decline.

Biologically speaking, the case $\gamma < \gamma^z_{\text{crit}}$ could correspond to a set of genes with complex regulation and a small coding sequence target for both nonfunctionalization and for the accumulation of synonymous mutations. This analysis predicts that genes with the features of complex regulation requiring multiple functional transcription factors that have the ability to be subfunctionalized together with a short coding sequence would be strong candidates for subfunctionalization rather than nonfunctionalization, and would be less likely to be characterised by a concave hazard function. These conditions could be met, for example, when genes are expressed in multiple tissues at different levels.

At the individual gene level, there are several classes of proteins that might be thought of as candidates for falling into this space. Casein is a longer protein that could accumulate synonymous changes, but would be hard to nonfunctionalize. It is expressed in multiple tissues, but the regulation of its expression and the strength of negative selection on each regulatory domain is not well known [132].

Another example of genes that might fall into this category are hormones like insulin and gonadotropin hormone releasing hormone (GnRH) that are relatively short proteins, although they are less broadly expressed [132]. GnRH has in fact been retained

after multiple gene duplication events in vertebrate lineages with functional divergence between copies (the functions in the ancestral state are not known) [104].

Our final example of genes that are candidates for this behaviour are the intrinsically disordered proteins, which are shorter than folded proteins on average, and may be more mutationally robust to nonfunctionalizing mutations [120]. What is unclear at this stage is the selection on their function and their expression.

While these types of genes are not likely to dominate any whole genome analysis (and hence, we would not conclude from our estimates that this is the dominant mode of subfunctionalization), the model predicts that genes with small mutational footprints for nonfunctionalizing mutations and large footprints for regulatory subfunctionalization would undergo subfunctionalization at high rates.

Two distinct avenues for extension of the work presented in this chapter are apparent. First, the model could be applied to a broader analysis. Here we applied our model in a whole-genome analysis to get some estimates of mutation rates and number of regulatory regions. The larger problems of the inference of parameters in a phylogenetic context, gene tree/species tree reconciliation, and ancestral copy number inference from multi-species data in a phylogenetic context are of particular interest. We will investigate the application of this model to ancestral copy number inference in future work.

The second avenue for extension, which we will discuss further in Chapter 4, is in widening the scope of the model itself. The model could be extended to include other processes contributing to the evolution of gene duplicates besides subfunctionalization, particularly neofunctionalization. Neofunctionalization can be conceptualised in a similar framework to subfunctionalization, with the same pathways to pseudogenization but where a new beneficial mutation in a regulatory region is the mechanism by which preservation of both duplicates can occur. We discuss how our existing subfunctionalization model could be modified to include neofunctionalization in Section 4.1. Other processes, such as of dosage balance, and the processes described in Innan and Kondrashov [55] could potentially also be implemented to create a more complete model for the evolution of gene duplicates. Another extension to the model which is of immediate interest is to account for larger gene families, which we will discuss in Section 4.2.

Chapter 4

Further Analysis of Duplicate Genes

In this chapter, we extend the model from Chapter 3 to allow for the analysis of the evolution of duplicate genes evolving with more complex dynamics, as detailed below.

First, in Section 4.1 we extend the model to include the process of neofunctionalization. Two models are presented for two different modes of neofunctionalization discussed in the literature, and we present some analysis of the second model. We fit the second model (Section 4.1.2) to the same dataset analysed in Section 3.8, and conclude that neofunctionalization is not a significant contributor to the preservation of gene duplicates over the timescales during which regulatory subfunctionalization occurs. We conclude that neofunctionalization is only likely in the presence of some other factor preserving the duplicates over the longer timescales necessary for neofunctionalization to occur with any significant probability.

The second model (Section 4.1.2) is introduced by first defining an initial, more intuitive model, and then mapping this model to a model with smaller state space. We apply the same algebraic procedure for model construction in Section 4.2, where the situation is more complex; the model development of Section 4.1.2 provides an example of this procedure in a simpler setting.

In Section 4.2 we extend the model from Chapter 3 to model the evolution of a family of gene duplicates (with a pair of gene duplicates being a family of size 2). We apply the algebraic procedure for model construction introduced by the example of the model in Section 4.1.2. In this way, we define the state space and generator of the model without explicitly considering the many possible transitions arising from the interdependence of the evolution of each gene in the family. We then consider the problem of computing the state space and generator explicitly, and outline an efficient procedure for doing so.

4.1 Modelling sub- and neofunctionalization for a pair of gene duplicates

In this section, we present a framework for the estimation of the relative importance of sub- and neofunctionalization for the preservation of gene duplicates, and some results from analysis of gene duplicate preservation data. Force [42] argued in his 1999 work that subfunctionalization was likely to be relatively more common than neofunctionalization, since it does not rely on beneficial mutation to preserve the pair of duplicates. It is generally thought that deleterious (and moreover, neutral) mutations are vastly more common than beneficial ones. Force posited that this could be investigated empirically by a biochemical investigation of the structure of duplicate pairs which have become fixed in genomes. Force noted that this would require recently fixed duplicates, hypothesising that subfunctionalization is likely to open the way for subsequent neofunctionalization in the remaining redundant regulatory regions after subfunctionalization has occurred. He and Zhang [50] performed one such analysis, investigating patterns of yeast protein interaction and human gene expression in duplicate genes. They found evidence for rapid subfunctionalization followed by neofunctionalization in a large proportion of duplicate genes.

We propose a different approach, using a mechanistically-motivated mathematical model based analysis built upon the ideas from the earlier sections of this chapter. We incorporate neofunctionalization alongside subfunctionalization in a model like the one introduced in Section 3.1, which we then fit to data similarly to Section 3.8. In this way, we are able to make some predictions about the relative likelihood of sub- and neofunctionalization. A potential issue with this approach would be the possibility of identifiability issues. As we have discussed in Sections 3.5, 3.8, and 3.9, subfunctionalization alone can produce the behaviour typically associated to neofunctionalization.

The term neofunctionalization is used to describe two similar but distinct biological models. The first is where one copy mutates to gain some new functionality at the expense of existing functionality, and the second is where one copy mutates to gain some new functionality without the loss of the existing functionality. To preserve the binary nature of the regulatory regions in our model, we can think of the second case as the addition of an extra regulatory region (a more biochemically accurate depiction is that an existing region gains new functionality). In this case it is possible to have neofunctionalization events which do not result in the preservation of both duplicates (when neofunctionalization occurs in the as-yet unmutated copy). The model in which new functionality is gained at the expense of existing functionality is simpler, since the loss of the original functionality ensures that both copies are preserved by selective

pressure whenever a neofunctionalization event occurs, and we treat this case first.

Note that it is not necessarily the case that a copy which has undergone neofunctionalization should always be protected by selective pressure. Clearly, the new function should not be assumed to be essential to the survival of the organism in the way that we have assumed existing functions are (since it was not there in the first place). Nonetheless, we make the simplifying assumption that neofunctionalization results in the copy being protected by selective pressure. Since we are primarily interested in measures pertaining to the time to the first neofunctionalizing event, it makes sense to treat it as such in the model. Further, it is possible that positive selection would act to protect the new function. A similar justification applies to our further assumption that each copy undergoes neofunctionalization at most once.

As we discussed in the Introduction, gene duplication is thought to contribute significantly to genome diversification by introducing additional (duplicate) genes, and hence, redundancy. Such redundancy could allow mutations which might otherwise be deleterious to fix and explore the space of possible sequences, creating space for beneficial changes to occur — this is in essence the motivation for the neofunctionalization model. Force [42] argued that subfunctionalization could amplify this effect by fixing a pair of gene duplicates with significant redundancy without requiring any beneficial mutation. This would increase the time over which the redundancy is preserved to allow for novel mutation, thereby increasing the chances of neofunctionalization. Thus, it is all the more pertinent that the two processes should be modelled together. To model this situation, we no longer treat subfunctionalization as absorbing, and instead introduce a sequence of transient states with 0 rate of absorption into P to track the post-subfunctionalization evolution of the sequence.

Remark 9.

Much of the model construction argument from Section 3.1 still applies, and we will not rehash it here — instead we will describe the new model in terms of the necessary modifications to the original one.

We will reuse the notation for the state space, generator, etc.

4.1.1 Model when neofunctionalization replaces functionality

Along with all of the states of the original model, we introduce the following additional states to model neofunctionalization, and its interaction with subfunctionalization:

- $2_S, ..., (z-1)_S$ — Transient state i_S corresponds to the situation in which i of the

regulatory regions across the two copies have had a null mutation fix, and both copies are uniquely associated with at least one regulatory region, but neither copy has undergone neofunctionalization;

- SN is an absorbing state corresponding to the situation in which subfunctionalization was followed by neofunctionalization; and
- N is an absorbing state corresponding to the situation in which neofunctionalization occurred in the absence of subfunctionalization.

Thus, the state space is

$$\mathcal{S} = \{0, ..., (z-1)\} \cup \{2_S, ..., (z-1)_S\} \cup \{S, SN, N, P\}. \tag{4.1}$$

The interpretation of absorbing state S here is slightly different to the model from Chapter 3. Previously, S corresponded to any case in which the pair had undergone subfunctionalization. We now track the evolution of the process after subfunctionalization, but it remains a possibility that the ultimate fate of the sequence is subfunctionalization alone. If none of the remaining regulatory regions are subject to nonfunctionalization, and subfunctionalization has occurred in the absence of neofunctionalization, then the process is absorbed into state S.

We define an ordering on $\{2_S, ..., (z-1)_S\}$ such that $i_S < j_S \iff i < j$. For the purpose of calculating transition rates, we treat the state i_S as the integer i, so that e.g. $(z - i_S)u_r = (z - i)u_r$.

Assuming that neofunctionalization occurs at Poisson rate u_n in each of the (still functional) regulatory regions of both copies, we infer the following transition rates:

- Transitions from $i \to (i+1)_S$ occur at rate $(z-i)u_r$ for $i = 1, ..., (z-2)$, and from $(z-1)$ to S at rate u_r — this is the equivalent situation to transitions from $i \to S$ in the original model, and the argument is the same.
- The remaining transition rates between the states included in the original model are unchanged.
- Transitions from $i_S \to (i+1)_S$ occur at rate $(z-i)u_r$ for $i = 1, ..., (z-2)$ and from $(z-1)_S$ to S at rate u_r — this is analogous to the transitions for states $1, ..., (z-1)$, but with no transitions to P.
- Transitions from $i \to N$ occur at rate $(z-i)u_n$ for $i = 1, ..., z-1$. Since we assume that neofunctionalization is associated with the loss of the original

functionality of the associated region, an unmutated copy of each regulatory region must be preserved, and hence there are $(z-i)$ susceptible regions, each undergoing neofunctionalization at rate u_n.

- Transitions from $i_S \to SN$ occur at rate $(z-i)u_n$ for $i = 1, ..., z-1$, by the same reasoning as the previous case.

Thus, we define the generator of our Markov chain to be $\mathbf{Q} = [q_{ij}]$ where the non-zero off-diagonals are given by

$$q_{ij} = \begin{cases} 2u_c & \text{if } i = 0, j = P \\ 2zu_r & \text{if } i = 0, j = 1 \\ 2zu_n & \text{if } i = 0, j = N \\ u_c & \text{if } 1 \leqslant i \leqslant z-2, j = P \\ (z-i)u_r & \text{if } 1 \leqslant i \leqslant z-2, j \in \{i+1, (i+1)_S\} \\ & \text{or } 2_S \leqslant i \leqslant (z-2)_S, j = (i+1)_S \\ (z-i)u_n & \text{if } 1 \leqslant i \leqslant z-2, j = N \\ & \text{or } 2_S \leqslant i \leqslant (z-2)_S, j = SN \\ u_r + u_c & \text{if } i = z-1, j = P \\ u_r & \text{if } i = z-1, j = S. \end{cases} \tag{4.2}$$

Note that, in terms of whether one or both copies are ultimately preserved, the main difference between this model and the original is that there is a non-zero transition rate from state 0 into N. Besides this, as far as the rate of absorption into state P is concerned the behaviour of this model is precisely the same as the original with u_r replaced by $u_r + u_n$. From this perspective, the model could be significantly simplified by combining the absorbing class $\{2_S, ..., (z-1)_S, S, SN, N\}$ into a single absorbing state (say, S) to represent both genes being preserved by selective pressure. Thus, in order to calculate the pseudogenization rate it is much more computationally efficient

to instead use $\mathbf{Q}^\dagger = [q^\dagger_{ij}]$ where,

$$q^\dagger_{ij} = \begin{cases} 2u_c & \text{if } i = 0, j = P \\ 2zu_r & \text{if } i = 0, j = 1 \\ 2zu_n & \text{if } i = 0, j = N \\ u_c & \text{if } 1 \leqslant i \leqslant z-2, j = P \\ (z-i)u_r & \text{if } 1 \leqslant i \leqslant z-2, j = i+1 \\ (z-i)(u_r+u_n) & \text{if } 1 \leqslant i \leqslant z-2, j = S \\ u_r + u_c & \text{if } i = z-1, j = P \\ u_r + u_n & \text{if } i = z-1, j = S. \end{cases} \tag{4.3}$$

However, as discussed above the interaction between subfunctionalization and neofunctionalization is itself of some interest — particularly the relative probability of absorption into N, S and SN. Moreover, the context in which neofunctionalization replaces functionality is usually one in which it is treated separately from subfunctionalization. In terms of modelling, the effect of a neofunctionalizing mutation replacing the functionality of a regulatory region is equivalent to having an initial nonfunctionalizing mutation followed by a neofunctionalizing mutation which does not replace functionality. To that end, we now construct and analyse a model for the situation in which neofunctionalization adds functionality without destroying any existing functionality.

4.1.2 Model when neofunctionalization adds functionality

In the case where neofunctionalization replaces the functionality of a regulatory region (discussed above) we introduced additional states to track the evolution of the duplicate pair after subfunctionalization. For the situation where neofunctionalization does not replace any existing functionality, it is possible to have neofunctionalization occur in the unmutated copy without leading to the preservation of both. In this case, it makes sense to track the evolution of the sequence post-neofunctionalization as well as post-subfunctionalization — certainly we should track the evolution after a neofunctionalization which does not lead to preservation of both copies. With this in mind, the simplest course is to track the evolution of the pair until all of the original regulatory regions are fully resolved (in the sense that none remain vulnerable to null mutation). We still assume that neofunctionalization occurs at most once in each copy.

To model this situation, we first introduce a more natural, and conceptually simple model, before defining a mapping from this model to a more computationally efficient one. We refer to the first model and its associated state space, etc. as *unreduced*, and the second as *reduced*. In this case, it would not be too challenging to skip the unreduced model and go straight to defining the reduced one, however this serves as a good first example of the procedure we apply for complex model development, which we will rely upon in Section 4.2. The procedure allows us to define an intuitively obvious model with many redundant states, and then apply algebra to define a minimal model in terms of the first one.

The usual approach to model development (applied e.g. in Section 3.1) explicitly indexes the state space, and relies on intuitions of the process being modelled to define the transition rates between each state. The procedure we apply is no different in theory, but somewhat more abstract, and very convenient for models with unwieldy state spaces.

To summarise the procedure, we first define the unreduced model in terms of sets of *transition functions* mapping between the states according to simple intuitions of the physical process. We then define a *rate function* mapping the transition functions to associated rates, easily intuited from the physical process. The generator is defined in terms of sums of the rate function over the transition functions. We avoid the need to explicitly consider the exact rate at which one state transitions to another. Instead, we rely only on very-simple intuitions of the physical process, with the more complex interactions being handled by the algebra. The state space of the reduced model is then defined in terms of a partition of the unreduced model's state space, and analogous mappings for the reduced model are defined in terms of the mappings of the unreduced model. This results in an optimised (in terms of the size of the state space) model without the need to carefully consider transition rates between states, which again, are handled by the algebra.

The approach is 'modular' in the sense that all of the distinct mechanisms of the physical processes are handled by their own set of functions, which lends itself extremely well to coding the model in MATLAB (or any other computing environment).

The unreduced state space is given by,

$$\mathcal{S}^{\ddagger} = \left\{ [s_{ij}] = \begin{bmatrix} r_1 & n_1 \\ r_2 & n_2 \end{bmatrix} : r_l \in \{0, 1, ..., z\}, r_1 + r_2 \leqslant z, n_l \in \{0, 1\} \right\}, \tag{4.4}$$

where r_1 and r_2 represent the number of regulatory regions which have undergone null mutation in the first and second (arbitrarily ordered) copy respectively, and n_1 and n_2 track neofunctionalization in the first and second copy respectively. The assumption

that a functioning copy of each region is protected implies that $r_1 + r_2 \leqslant z$.

Any state for which $r_1 + r_2 = z$ is absorbing. States with $r_l = z$ and $n_l = 0$ correspond to pseudogenization, and those for which $r_1 > 0$ and $r_2 > 0$ correspond to subfunctionalization (which is not necessarily absorbing).

To simplify the transition discussion, we define the following propositions.

Definition 57 (Subfunctionalization proposition)**.**

$S(s)$ is the proposition

$$s_{i1} > 0 \text{ for } i = 1, 2,$$

and $\overline{S}(s)$ is its complement.

Definition 58 (Absorption proposition)**.**

$A(s)$ is the proposition

$$s_{11} + s_{21} = z,$$

and $\overline{A}(s)$ is its complement.

Definition 59 (Neofunctionalization proposition)**.**

$N(s)$ is the proposition

$$s_{12} = 1 \text{ or } s_{22} = 1,$$

and $\overline{N}(s)$ is its complement.

We define transition functions corresponding to each of the kinds of mutations which can occur in the process, carefully excluding states from the domains of each for which there are no susceptible regions to undergo the mutation.

Definition 60 (Unreduced regulatory mutation function)**.**

$\mathcal{R}_l^\ddagger$ is a function on $\{s \in \mathcal{S}^\ddagger : \overline{A}(s)\}$ such that

$$\mathcal{R}_l^\ddagger(s)_{ij} = \begin{cases} s_{ij} + 1 & \text{for } i = l, j = 1 \\ s_{ij} & \text{otherwise.} \end{cases} \tag{4.5}$$

Definition 61 (Unreduced coding mutation function)**.**

$\mathcal{C}_l^\ddagger$ is a function on $\{s \in \mathcal{S}^\ddagger : [s]_{k1} = 0, k \neq l$ and $[s]_{l2} = 0$ and $\overline{S}(s)$ and $\overline{A}(s)\}$ such that

$$\mathcal{C}_l^\ddagger(s)_{ij} = \begin{cases} z & \text{for } i = l, j = 1 \\ s_{ij} & \text{otherwise.} \end{cases} \tag{4.6}$$

Definition 62 (Unreduced neofunctionalization mutation function)**.**

$\mathcal{N}_l^\ddagger$ *is a function on* $\{s \in \mathcal{S}^\ddagger : [s]_{l2} = 0$ *and* $\overline{A}(s)\}$ *such that*

$$\mathcal{N}_l^\ddagger(s)_{ij} = \begin{cases} 1 & \text{if } i = l, j = 2 \\ s_{ij} & \text{otherwise.} \end{cases} \tag{4.7}$$

Together, these functions account for all possible transitions of the combined neo- and subfunctionalization process, and we can associate with each function the following rates for $i = 1, 2$:

- $\mathcal{R}_i^\ddagger(s)$ is associated with a rate $(z - s_{11} - s_{21})u_r$ — since s_{11} of the i^{th} copies' regulatory regions are either already nonfunctionalized, while s_{21} are protected by selective pressure (being nonfunctionalized in the other copy), leaving $(z - s_{11} - s_{21})$ vulnerable to nonfunctionalization, which they each do at rate u_r;
- $\mathcal{C}_i^\ddagger(s)$ is associated with a rate u_c — since the existence of $\mathcal{C}_i^\ddagger(s)$ ensures that the i^{th} copy is susceptible to nonfunctionalization of its coding region (leading to pseudogenization) at rate u_c; and
- $\mathcal{N}_i^\ddagger(s)$ is associated with a rate $(z - s_{i1})$ — since the i^{th} copy has $(z - s_{i1})$ regulatory regions remaining to undergo neofunctionalization, which they each do at rate u_n.

The transition rate from $s \to x$ can then be found by summing the rates associated with each function mapping s onto x. Examining the functions we can see that the only instance where any two of these functions map from and to the same state occurs for $\mathcal{R}_i^\ddagger$ and $\mathcal{C}_i^\ddagger$. This can be seen by noting the fact that besides these two pairs (with $i = 1$ or 2), no other pair act on a the same matrix entry as each other. Moreover, $\mathcal{R}_i^\ddagger(s) = \mathcal{C}_i^\ddagger(s)$ if and only if $s_{i1} = z - 1$. Note that this is in keeping with our earlier analysis in Section 3.1, from which we already know that this transition occurs at a rate $u_c + u_r$, as confirmed by the sum of the rates associated with $\mathcal{R}_i^\ddagger(s)$ and $\mathcal{C}_i^\ddagger(s)$. Accounting for this, we can use these functions to define our generator.

Further, to index the states we define an arbitrary bijection $f^\ddagger$ from $\mathcal{S}^\ddagger \to \{1, ..., \left|\left|\mathcal{S}^\ddagger\right|\right|\}$. We will refer to both state s and its mapping under the bijection $f^\ddagger(s)$ by s, with the understanding that whenever s appears in an index, we are referring to $f^\ddagger(s)$, and otherwise, to state s.

The final part of the setup is to define $\mathcal{F}^\ddagger$ to be the set of all the previously introduced functions, together with a function $r^\ddagger(f, s)$ from $\mathcal{F}^\ddagger \times \mathcal{S}^\ddagger$ (being a Cartesian product)

to $\mathbb{R}$ such that, for example $r^{\ddagger}(\mathcal{R}_i, s) = (z - s_{11} - s_{21})u_r$ — the definition is analogous for the other functions, with rates given as discussed in the bullet points above.

Then we define the generator for our Markov chain to be $\mathbb{Q}^{\ddagger} = q^{\ddagger}_{sx}$ where the nonzero off-diagonals are given by

$$q^{\ddagger}_{sx} = \sum_{\{f \in \mathcal{F}^{\ddagger} : f(s) = x\}} r^{\ddagger}(f, s), \tag{4.8}$$

or, equivalently,

$$q^{\ddagger}_{sx} = \begin{cases} u_c & \text{if } x = \mathcal{C}^{\ddagger}_i(s) \neq \mathcal{R}^{\ddagger}_i(s) \\ u_c + u_r & \text{if } x = \mathcal{C}^{\ddagger}_i(s) = \mathcal{R}^{\ddagger}_i(s) \\ (z - s_{11} - s_{21})u_r & \text{if } x = \mathcal{R}^{\ddagger}_i(s) \neq \mathcal{C}^{\ddagger}_i(s) \\ (z - s_{i1})u_n & \text{if } x = \mathcal{N}^{\ddagger}_i(s), \end{cases} \tag{4.9}$$

with $i = 1, 2$.

However, as mentioned there is some redundancy in the state space here, since any two states which are equivalent up to row swapping are equivalent for our modelling purposes (with the two copies being only arbitrarily indexed as the first and second copy). Of course, it is optimal to minimize the size of the state space for computational purposes, and thus we will present a version of the model with a minimal 'reduced' state space. To that end, we define the following relation on $\mathcal{S}^{\ddagger}$,

Definition 63 (Row-equivalence)**.**

Define binary relation $\leftrightarrow$ on $\mathcal{S}^{\ddagger}$ by $s \leftrightarrow x$ if either $s = x$ or swapping the rows of s yields x. We say s and x are row equivalent.

Clearly $\leftrightarrow$ is an equivalence relation.

Definition 64 (Reduced State space $\mathcal{S}$)**.**

The reduced state space $\mathcal{S}$ is defined to be the quotient set of $\mathcal{S}^{\ddagger}$ by $\leftrightarrow$. Each equivalence class is represented by its member with the largest first row sum, or in the case of equal row sums, the largest first entry, i.e. s such that, if $s \leftrightarrow x$ then,

$$s_{11} + s_{12} \geqslant x_{11} + x_{12}, \tag{4.10}$$

and, in the case of equality in (4.10),

$$s_{11} \geqslant x_{11}, \tag{4.11}$$

If the equality holds in both (4.10) and (4.11), then $s = x$. This can be seen by noting that $s_{1i} = x_{2i}$, so the two equalities hold only if the rows of s are equal, and hence $s = x$. Thus the representative is unique, as required. This associates the first row of representative state s with the copy which has undergone the most mutations, and in the case of a tie, the most deleterious (i.e. not neofunctionalizing) mutations.

Definition 65 (Order of s)**.**

We define the order of state $s \in \mathcal{S}$, denoted $|s|$ to be the number of elements in its equivalence class under $\leftrightarrow$, i.e.

$$|s| = ||\{x \in S^* : x \leftrightarrow s\}|| \tag{4.12}$$

Clearly $|s| = 1$ or $|s| = 2$ for all $s \in \mathcal{S}$, since either the rows of s are equivalent, or they are not.

We define functions on $\mathcal{S}$ as follows

Definition 66. *(Reduced regulatory mutation function)*

$\mathcal{R}_l$ is a function on $\mathcal{S}$ defined by $\mathcal{R}_l(s) = x$ such that $x \leftrightarrow \mathcal{R}_l^{\ddagger}(s), x \in \mathcal{S}$.

Definition 67. *(Reduced coding mutation function)*

$\mathcal{C}_l$ is a function on $\mathcal{S}$ defined by $\mathcal{C}_l(s) = x$ such that $x \leftrightarrow \mathcal{C}_l^{\ddagger}(s), x \in \mathcal{S}$.

Definition 68. *(Reduced neofunctionalization mutation function)*

$\mathcal{N}_l$ is a function on $\mathcal{S}$ defined by $\mathcal{N}_l(s) = x$ such that $x \leftrightarrow \mathcal{N}_l^{\ddagger}(s), x \in \mathcal{S}$.

We define $\mathcal{F}$ and $r(f, s)$ analogously to $\mathcal{F}^{\ddagger}$ and $r^{\ddagger}(f, s)$, together with a bijection in the same manner as we did previously.

The rates associated with the reduced functions are the same as they were for the unreduced functions, and like the unreduced functions, $\mathcal{R}_i(s) = \mathcal{C}_i(s)$ whenever $s_{i1} = z - 1$. However whenever $|s| = 1$, we also have $\mathcal{R}_1(s) = \mathcal{R}_2(s)$, $\mathcal{C}_1(s) = \mathcal{C}_2(s)$ and $\mathcal{N}_1(s) = \mathcal{N}_2(s)$. Then the generator for our Markov chain is given by $\mathbb{Q} = q_{sx}$, with non-zero off diagonals given by

$$q_{sx} = \sum_{\{f \in \mathcal{F} : f(s) = x\}} r(f, s), \tag{4.13}$$

or, equivalently,

$$q_{sx} = \begin{cases} 2(z-2s_{11})u_r & \text{if } |s|=1, x=\mathcal{R}_1(s) \\ 2(z-s_{11})u_n & \text{if } |s|=1, x=\mathcal{N}_1(s) \\ 2u_c & \text{if } |s|=1, x=\mathcal{C}_1(s) \\ u_c & \text{if } |s|=2, x=\mathcal{C}_i(s)\neq\mathcal{R}_i(s) \\ u_c+u_r & \text{if } |s|=2, x=\mathcal{C}_i(s)=\mathcal{R}_i(s) \\ (z-s_{11}-s_{21})u_r & \text{if } |s|=2, x=\mathcal{R}_i(s)\neq\mathcal{C}_i(s) \\ (z-s_{i1})u_n & \text{if } |s|=2, x=\mathcal{N}_i(s), \end{cases} \tag{4.14}$$

with $i=1,2$. Note that when $|s|=1$ we have $s_{11}=s_{21}$, hence the simplification of the first line. Also, it is never the case that $s \leftrightarrow x$ with $x=\mathcal{R}_i(s)=\mathcal{C}_i(s)$, since if $s_{11}=s_{12}=0$ then $\mathcal{R}_i(s)\neq\mathcal{C}_i(s)$, and otherwise $\mathcal{C}_i(s)$ is not defined for s with $s_{11}=s_{12}$. Thus, this case is not included.

Naturally, we would expect the subfunctionalization model (Section 3.1) and the sub- and neofunctionalization model to be equivalent when the rate of neofunctionalization is zero. Thus we tested consistency of the two models for $u_n=0$.

We consider the generator of the sub- and neofunctionalization model with the rows and columns corresponding to the states which are inaccessible from s_0 when $u_n=0$ (i.e. the ones associated with neofunctionalization) removed. Also, states associated subfunctionalization are treated as a single absorbing state S (since the subfunctionalization model treats it as such). Thus the transitions associated with $\mathcal{N}_i(s)$ are removed, and any transitions leading to s with $s_{21}\neq 0$ are redirected to S.

In this case, the generators of the two models are equivalent up to the indexing of the states. This can be seen by carefully examining Equation (4.13), which reduces to Equation (3.2) with different notation,

$$q_{sx}^{u_n=0} = \begin{cases} 2u_c & \text{if } s=s_0, x_{11}=z \\ 2zu_r & \text{if } s=s_0, x_{11}=1 \\ u_c & \text{if } 1\leqslant s_{11}\leqslant z-2, x_{11}=s_{11}+1 \\ (z-s_{11})u_r & \text{if } s_{11}=z-1, x=s_{11}+1 \text{ or } x=S \\ u_c+u_r & \text{if } s_{11}=z-1, x_{11}=z \\ u_r & \text{if } s_{11}=z-1, x=S. \end{cases} \tag{4.15}$$

The pseudogenization rate function for the sub- and neofunctionalization model can

be defined analogously to Equation (3.34),

$$h(t) = \frac{\underline{e_0} e^{\mathbf{Q}^* t} \boldsymbol{V}_P \underline{1}}{1 - \underline{e_0} \left(e^{\mathbf{Q}^* t} - \mathbf{I}\right) \left(\mathbf{Q}^*\right)^{-1} \boldsymbol{V}_P \underline{1}}, \tag{4.16}$$

where, $\boldsymbol{V}_P$ is a $* \times 2$ matrix of transition rates from the transient states into P, where $P = \{s_p, s_{pn}\}$ is the collection of states which correspond to pseudogenization (there are only ever two in the reduced state space),

$$s_p = \begin{bmatrix} z & 0 \\ 0 & 0 \end{bmatrix}, \text{ and } s_{pn} = \begin{bmatrix} z & 0 \\ 0 & 1 \end{bmatrix}, \tag{4.17}$$

with s_p corresponding to the usual pseudogenization scenario, and s_{pn} corresponding to the case where the preserved copy underwent neofunctionalization before the pseudogenization of the other copy.

Further, the extension of the model to a population of gene duplicates evolving under sub- and neofunctionalization is exactly analogous to the analysis in Section 3.7. The log likelihood given by Equations (3.107) and (3.108), but with $\tilde{F}(t)$ taken as the cumulative distribution of time to absorption into the states associated with pseudogenization from this model, i.e. the set $\{s_p, s_{pn}\}$. We restate these here for convenience,

$$\beta(t) = \beta_0(1 - \tilde{F}(t)), \tag{3.107}$$

$$\log(L_\theta|D) = \sum_i D_i \log(\beta(s_i)) - \beta(s_i) - \Gamma \log(D_i + 1). \tag{3.108}$$

4.1.3 Results

In this section, we discuss the data-driven analysis of the model described in Section 4.1.2. We have fit the model to the *Mus musculus* and *Homo Sapiens* data from Section 3.8. We excluded the *Canis Familiaris* and *Rattus norvegicus* datasets on the basis of the earlier analysis in Section 3.8, which showed these genomes to be uninformative. Further, we have tested the effect of including neofunctionalization on the pseudogenization rate as compared to the subfunctionalization model in Section 3.1. We examine a wide range of u_n with the remaining parameters fixed at biologically realistic values. We calculate the probability of neofunctionalization, and the relative probability of neofunctionalization before and after subfunctionalization given that it occurs at all.

Fitting to the *Mus musculus* and *Homo Sapiens* genomes gave similar parameter estimates (for the shared parameters) to the earlier model (with MLEs given in Table 3.1 of Section 3.8). The maximum likelihood parameter estimates of u_r and u_c for the

two models were identical to at least the first 5 significant figures for both genomes. The MLE for the neofunctionalization rate was $u_n \approx 10^{-10}$ for both. u_n was similarly small for all but $z = 2$, for which $u_n = 2.52 > u_r = 1.74$, and $u_n = 2.37 > u_r = 1.60$ for *Mus musculus* and *Homo sapiens* respectively. The likelihood values themselves were also essentially-equal to the subfunctionalization-only model (Section 3.8), which would be preferred by either BIC or AIC (see Section 2.2).

As mentioned in Section 4.1, we expected that there may be identifiability issues in fitting this model to data. Our reasoning was based on the conclusion from Chapter 3 that subfunctionalization alone could lead to survival distributions of the kind which have previously been attributed to neofunctionalization. We did not find model identifiability to be a problem in practice, with a range of starting points converging to the same maximum likelihood parameter estimates for u_r, u_c, and z. However there was some variation in u_n. Moreover, for both genomes, the difference in likelihood between $u_n = 10^{-10}$ and $u_n = 10^{-3}$ (with the other parameters at their MLE values) was extremely small. The two likelihood values were equal up to the 6^{th} and 5^{th} significant figure for *Mus Musculus* and *Homo Sapiens* respectively. So, although we were able to identify a clear maximum likelihood estimate, the parameter u_n in particular would be associated with a wide confidence interval.

Our estimates were based on intervals with $s = 0.01$, corresponding to roughly 1.1 million years (as discussed in Section 3.8). As such, $u_n \approx 10^{-10}$ implies that on average 1 out of every 10^{10} regulatory regulatory regions would neofunctionalize every 110 million years. With roughly 20000 genes in the genomes we examined, even assuming all of these genes are targets for neofunctionalization at all times, with 3 regulatory regions each $u_n \approx 10^{-10}$ would correspond to the fixation of less than one beneficial mutation on average over the entire history of life on earth. Clearly, $u_n = 10^{-10}$ is an underestimate. Since the likelihood did not drop off significantly until $u_n > 10^{-3}$, estimates closer to this end of the interval seem more realistic. In either case, our results are suggestive of a very low rate of neofunctionalization as compared to regulatory nonfunctionalization.

We computed $h(t)$ (Equation (3.34)) and $h_n(t)$ (Equation (4.16) above) at the maximum likelihood parameter estimates of $u_c = 20.1, u_r = 3.26$, and $z = 3$ from the *Mus musculus* genome for a range of u_n. The rate for $h(t)$, and $h_n(t)$ with $u_n = 0.001u_r, 0.01u_r$ and $0.1u_r$ is shown in Figure 4.1. The rates were somewhat divergent for $u_n = 0.1u_r$, but a ratio of $1 : 10$ beneficial to neutral mutations in the regulatory regions is extremely high, and it vastly exceeds the maximum likelihood estimate of 4.8×10^{-10}. For the other values, the graphs were very close in a by-

eye examination, with $u_n = 0.001$ indistinguishable by-eye from $u_n = 0$ (noting that $h(t) = h_n(t)$ when $u_n = 0$).

We calculated the probability of neofunctionalization before and after subfunctionalization (which is of particular interest). We treated the states associated with neofunctionalization as absorbing, and computed the probability that the process (modified such that neofunctionalization is absorbing) is eventually absorbed into states associated with neofunctionalization only, and with neofunctionalization following subfunctionalization. We chose u_r and u_c as the MLE parameters from the *Mus musculus* genome again. Figure 4.2 shows the probability that the process ever neofunctionalizes as as a function of u_n in the range from 0 to $0.5u_r$ for $z = 3, ..., 10$. Figure 4.3 shows the probability of neofunctionalization before subfunctionalization conditional on neofunctionalization occuring as a function of u_n for $z = 3, ..., 10$, while Figure 4.2 shows the probability of neofunctionalization as a function of u_n (of course, this goes to 0 as u_n does). Most likely u_n is orders of magnitude smaller than u_r, since beneficial mutations are expected to be very rare [42], and our fit to the *Mus musculus* and *Homo Sapiens* genomes supports this.

Based on this analysis, it appears that neofunctionalization is not a significant contributor to the preservation of gene duplicates. Neofunctionalizing mutations appear to be extremely unlikely during the timescales over which regulatory subfunctionalization is resolved. Thus copies which do not undergo subfunctionalization are lost before neofunctionalization has a chance to occur. Given that neofunctionalization does occur, the probability (associated with the MLEs) that it occurs before subfunctionalization was 0.80 for both genomes.

From the modelling perspective, this is an intuitive result, given that for $z = 3$, after subfunctionalization occurs the next non-neofuntionalizing mutation must lead to the absorption of the process. As such, the window of opportunity for neofunctionalization after subfunctionalization is likely to be much shorter than it is beforehand in this case. However, this result only applies for the timescales over which regulatory subfunctionalization is resolved, and should not be regarded as counter evidence to the hypothesis that subfunctionalization plays a protective role in support of subsequent neofunctionalization.

In reality, it is likely that neofunctionalization may occur over much larger timescales — potentially, the non-functionalized regulatory regions would have a non-zero rate of neofunctionalizing mutations, provided that the coding region remains intact. A model in which the nonfunctionalized regulatory regions can become neofunctionalized at some small rate may be more realistic in this sense, and the proportion of neofunc-

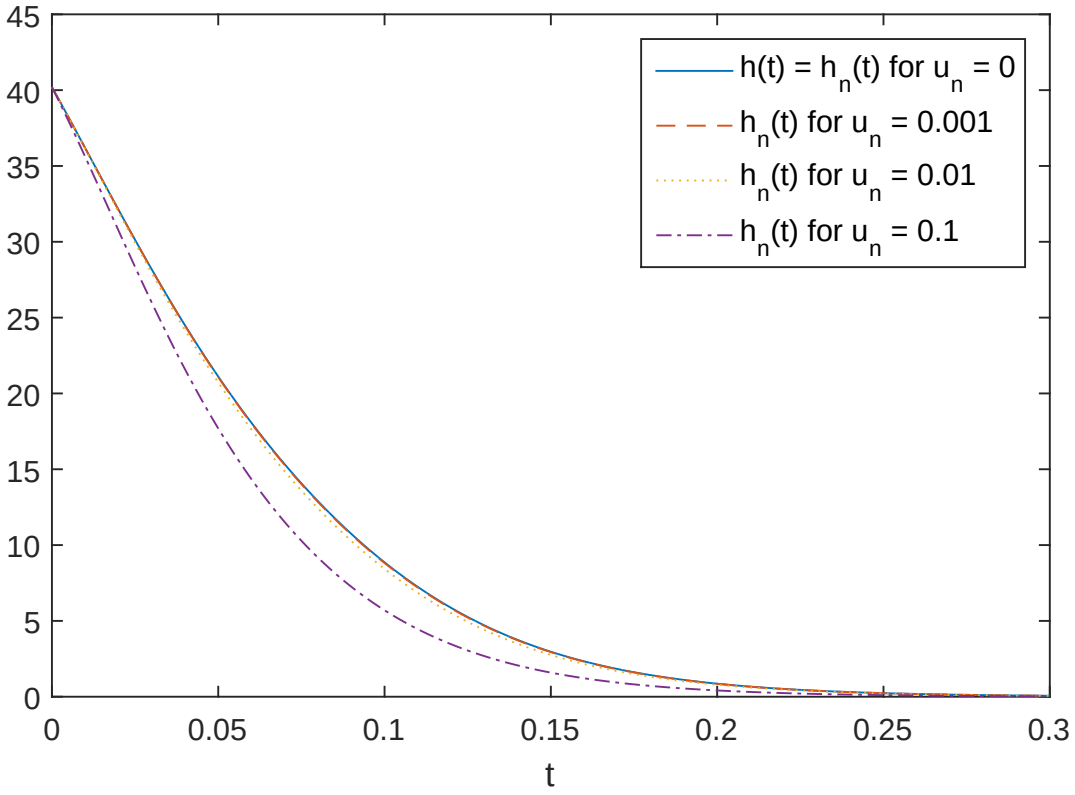

Figure 4.1: Pseudogenization rate $h_n(t)$ for $u_n = 0,\ 0.001u_r,\ 0.01u_r, 0.1u_r$ with $u_c = 20.1$, $u_r = 3.26$ and $z = 3$, which were the maximum likelihood parameter estimates associated with the *Mus musculus* genome. $h(t)$ is given by Equation (3.34), while $h_n(t)$ is given by Equation (4.16).

tionalization events preceded by subfunctionalization could only increase under such assumption.

Rastogi and Liberles [97] concluded from a lattice model analysis that neofunctionalization was the ultimate fate for all preserved gene duplicates. They noted that subfunctionalization could act to protect the genes from pseudogenization during short timescales, and our results appear to support that hypothesis. Under our model, if a gene is preserved at all, it is almost certainly due to sub- and not neofunctionalization. Subsequent neofunctionalization is extremely unlikely in the event that one of the copies becomes pseudogenized by null mutation in the coding region. Thus, for neofunctionalization to occur some other process must almost-always have preceded it to preserve the duplicate pair, and subfunctionalization appears to be a likely candidate.

In this regard, our analysis can be thought of as providing an estimate of a lower bound on the proportion of neofunctionalization events preceded by subfunctionaliza-

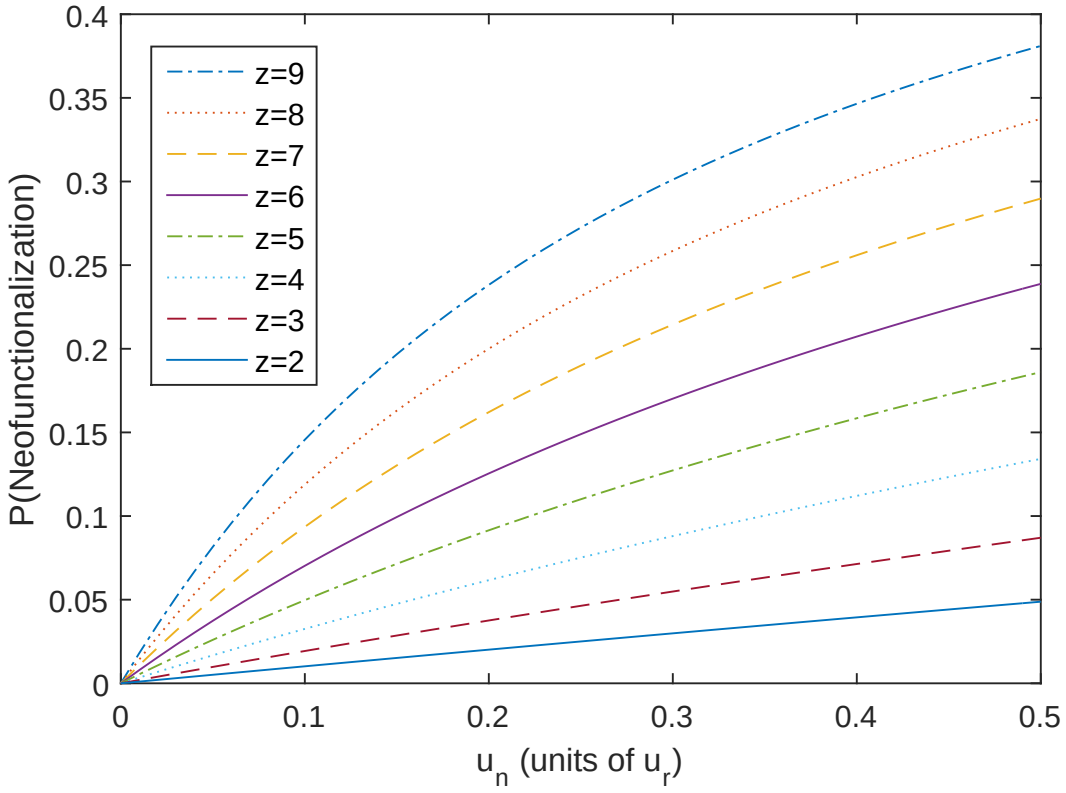

Figure 4.2: The overall probability of neofunctionalization as a function of u_n for $z = 2, ..., 9$, (plots are strictly ascending in z) with $u_c = 20.07, u_r = 3.26$.

tion (which we estimate to be 0.2 for $z = 3$, 0.35 for $z = 4$, 0.46 for $z = 5$ for the *Mus Musculus* genome). Given the extremely low probability of neofunctionalization during the timescale of our model, and the possibility for subsequent neofunctionalization over a much larger timescale, the proportion of neofunctionalization events preceded by subfunctionalization is likely much larger. A caveat is that processes other than subfunctionalization (most notably dosage balance [116]) could play a similar role, but based on this analysis neofunctionalization alone is not very likely.

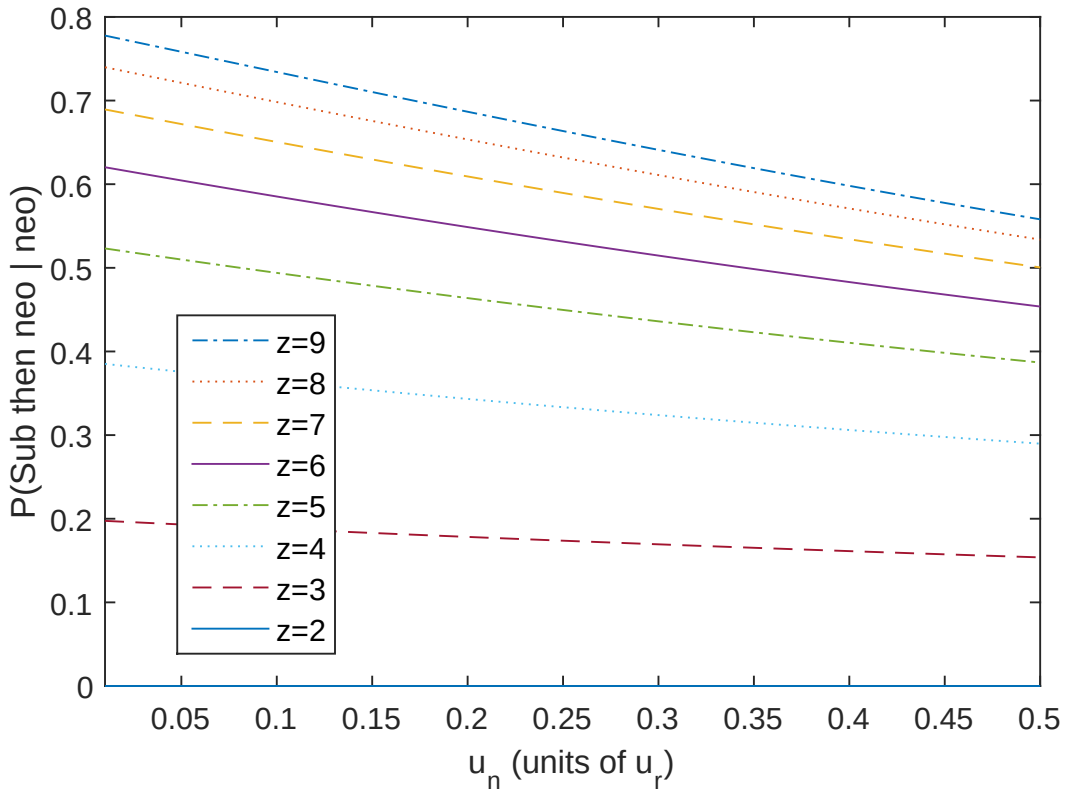

Figure 4.3: Probability that subfunctionalization occurs before neofunctionalization, conditional on neofunctionalization occurring as a function of u_n for $z = 2, ..., 9$, (plots are strictly ascending in z) with $u_c = 20.07, u_r = 3.26$. When $z = 2$ only one of sub- or neofunctionalization can occur under the model.

4.2 Preliminary work modeling the evolution of gene families

In this section, we extend the model introduced in Chapter 3 from the case in which there are two perfect copies of a particular gene to the case where there are n genes comprising a gene family. We start with an intuitive discussion of the approach we have in mind to model this situation, which is followed by a more rigorously defined model using the procedure introduced in Section 4.1.2. We then discuss some computational considerations of working with the model. This work is 'preliminary' in the sense that we only construct the model and discuss computation, no analysis of the model has been undertaken as yet. Further, the model introduced here is for the case of a gene family of fixed size; ultimately, we would like to extend this model to allow for additional duplication events to occur, and the size of the gene family to vary.

A gene family is a collection of genes all having arisen from duplication (or speciation) events with some common ancestor. Gene families comprised of many different numbers of genes have been observed in a range of extant genomes, and have been found to diverge significantly along lineages. For example, Demuth et al. [32] analysed the gene families of several mammalian species and found that "more than half of the 9,990 families present in the mammalian common ancestor have either expanded or contracted along at least one lineage." In their analysis of a soybean genome, Nelson and Shoemaker [86] found that 95% of gene families had 10 or fewer members. While much larger gene families (including so-called superfamilies) are observed, we will be interested primarily in modelling small families (i.e. up to 10 or-so members) to keep things relatively tractable.

Consider a collection of n genes, all having arisen from duplication events with some common ancestor gene. We assume that the z mutable regulatory regions and corresponding functions that were associated with the ancestral gene are distributed among the n gene family such that each function is present in at least one gene — each of the n genes can have a functioning or non-functioning copy of each of the z regulatory regions. Just as in the case with 2 copies (which can be thought of as a gene family with 2 members), we assume that null mutations fix in the regulatory and coding regions of each gene at Poisson rate u_r and u_c respectively.

For a fixed number z of regulatory regions, an initial model which comes to mind takes as its state space the $* \times z$ binary matrices, where each row represents a gene, each column represents a regulatory region, and a 1 in the (i, j) entry corresponds to gene i having a working copy of regulatory region j.

Then duplication events can be modelled by the insertion of an extra copy of an existing row, say at Poisson rate u_d for each existing gene, a null mutation in the coding region can be modelled by row deletion, and null mutations in the regulatory region can be modelled by replacing a 1 entry with a 0. Any row consisting entirely of 0s could be deleted. While such a model may not appear on first pass to be particularly tractable, ensuring a higher rate of pseudogenization than duplication should allow truncation of the state space, perhaps in the vein of matrix analytic methods [72], or in a data-inspired method like the truncation we apply in Chapter 5. At the very least, implementing a simulation via competing Poisson processes is straight-forward.

For such a process it may not be immediately obvious that any state should be absorbing, since new duplication events can always occur. However, the Markov model constructed based on these ideas is reducible, and a class of essentially-similar (from the biological perspective) final states exists. Intuitively we would expect that given enough time the process will eventually reach a state in which all of the row sums are 1 (since rows consisting entirely of 0's are deleted, and the row sum can only decrease). This property will be preserved from then on; in fact, no state with a higher maximum row sum than i is accessible from i, since there is no process by which a 0 entry will be replaced by a 1. Since further evolution is of little interest once each gene is associated with just one regulatory region, it makes sense to simplify the model such that any of these states is absorbing. Alternatively, neofunctionalization could be included in the model by the insertion of a column with sum 1, in which case the reducing row sum property clearly would not hold, and the process would be modelled without any absorbing states.

If the size of the gene family is fixed equal to n, and duplication events not allowed, then we take the $n \times z$ binary matrices as the state space and treat coding mutations as replacing an entire rows by 0's, rather than deleting it. In this case, we might treat any state for which each row not-entirely composed of zeros has an entry which is the only 1 in its column as absorbing, as this is the condition for all genes in the family to be protected by selective pressure under the subfunctionalization model. To continue to model the process after subfunctionalization has occurred (e.g. to include neofunctionalization), either those states for which all column sums are equal to one (hence every regulatory region is represented in exactly one gene) could be treated as absorbing.

Although the model need not necessarily be absorbing, the motivation for it it to be so is that the biological process we are interested in is one of short (in evolutionary time) bursts of evolution taking place during a period of redundancy created by an initial

duplication event before eventual fixation. This is followed by long periods of relatively little change. This is particularly relevant when considering subfunctionalization alone, though perhaps less so when modelling the combined sub- and neofunctionalization process, due to the longer timescales over which neofunctionalization occurs (as discussed in Section 4.1).

While this provides a sufficient framework for simulating the evolution of gene families of fixed or dynamic size n, the complex state space does not make for a particularly analytically tractable model. In either case, the naive model with the $* \times z$ or the $n \times z$ binary matrices has a lot of redundancy, and examining the state space algebraically will simplify the analysis somewhat. Since each of the genes and each of the regulatory regions is treated as being equivalent, it is clear that many of the states in the (implicit) models are equivalent from the biological perspective. In fact if state i can be transformed into state j by any sequence of row or column swaps (or sequences involving both), then i and j are biologically equivalent. This allows for the state space to be significantly reduced by instead considering only the equivalence classes under such swaps. We follow a similar model development procedure to that which we applied to construct the neofunctionalization model in Section 4.1.2.

4.2.1 Model for gene families of fixed size n

With the intuitions outlined, we now move onto a more rigorous discussion, we will be redefining some notation from earlier sections with the understanding that this section constitutes a separate analysis.

Let $\mathcal{B}^*$ be the $n \times z$ binary matrices — i.e. the set of $n \times z$ matrices whose entries are either 0 or 1. Clearly $||\mathcal{B}^*|| = 2^{nz}$. There is an obvious bijection between $\mathcal{B}^*$ and the binary representation of the non-negative integers up to $2^{nz} - 1$; with leading 0's included (up to nz digits), the first z digits form the first row, the next z the second row, etc.

We consider the evolution of a gene family, with a fixed number z of regulatory regions in each gene, and for which the mode of evolution is regulatory subfunctionalization alone. Consider a continuous-time Markov chain $\{X(t), t \geqslant 0\}$ with state space

$$\mathcal{S}^* = \{s = [s_{ij}] : s \in \mathcal{B}^*, \sum_{j=1}^{z} s_{ij} \geqslant 1 \, \forall \, i = 1, ..., n\}, \tag{4.18}$$

where $s_{ij} = 1$ if the i^{th} gene has a functioning copy of the j^{th} regulatory region, and $s_{ij} = 0$ if it has a nonfunctioning copy, or a mutation in the coding region has led to pseudogenization. Note that in the case where $n = 2$ it is not necessary to track

which copy particular mutations occur in, since in this case as soon as both copies have a mutation fix, the process is absorbed. In the $n > 2$ case, the dynamics of the process depend on which gene undergoes mutation, and thus we introduce the binary matrices to track this.

States for which $\sum_{i,j} s_{ij} = z$ are considered absorbing, since they correspond to each subfunction being uniquely associated to one gene (the ultimate fate of a family of genes under the DDC process [42]). Analogously to the construction in Section 4.1, we define some propositions and functions operating on the state space $\mathcal{S}$ to simplify the discussion.

Definition 69 (Subfunctionalization proposition 1)**.**

Let $S_i(s)$ be the proposition

$$s_{ij} = 1 \text{ for some } j \text{ such that } s_{kj} = 0 \text{ for all } k \neq j, \tag{4.19}$$

and $\overline{S}_i(s)$ be its complement.

Proposition $S_i(s)$ is true precisely when gene i is protected by selective pressure, since it is uniquely associated to some function.

Definition 70 (Subfunctionalization proposition 2)**.**

Let $S_{ij}(s)$ be the proposition

$$s_{ij} = 1 \text{ such that } s_{kj} = 0 \text{ for all } k \neq j, \tag{4.20}$$

and $\overline{S}_{ij}(s)$ be its complement.

Proposition $S_{ij}(s)$ is true precisely when regulatory region j of gene i is protected by selective pressure. Note that $S_{ij}(s) \implies S_i(s)$ for all j.

To track the process until regulatory subfunctionalization is fully resolved, we introduce the following proposition.

Definition 71 (Absorption proposition)**.**

Let $A(s)$ be the proposition

$$\sum_{i,j} s_{ij} = z, \tag{4.21}$$

and $\overline{A}(s)$ be its complement.

Often we are interested in tracking the process only until the number of preserved duplicates which will ultimately be preserved is resolved. In this case, we can replace the absorption proposition by the following, and reduce the state space accordingly.

Definition 72 (Alternative absorption proposition)**.**

Let $A^2(s)$ *be the proposition*

$$\left(\sum_j s_{ij} \neq 0 \implies \text{there exists } j \text{ with } s_{ij} = 1, \text{ and } \sum_k s_{kj} = 1\right) \text{ or } A(s). \tag{4.22}$$

We now define transition functions to handle regulatory and coding mutations, excluding those mappings which would correspond to biological impossibility.

Definition 73 (Regulatory-mutation function)**.**

Function $\mathcal{R}^*_{ij}$ *has domain* $\mathcal{S}_{\mathcal{R}_{ij}} = \{s \in \mathcal{S}^* : \overline{A}(s) \text{ and } \overline{S}_{ij}(s)\}$ *and is defined by, for all* $s \in \mathcal{S}_{\mathcal{R}_{ij}}$*,*

$$\mathcal{R}^*_{ij}([s_{kl}]) = [s_{kl} 1((k,l) \neq (i,j))], \tag{4.23}$$

where $1(.)$ *is an indicator function.*

Definition 74 (Coding-mutation function)**.**

We define function $\mathcal{C}^*_i$ *on* $\mathcal{S}_{\mathcal{C}_i} = \{s \in \mathcal{S}^* : \overline{A}(s) \text{ and } \overline{S}_i(s)\}$ *by, for all* $s \in \mathcal{S}_{\mathcal{C}_i}$*,*

$$\mathcal{C}^*_i([s_{kl}]) = [s_{kl} 1(k \neq i)]. \tag{4.24}$$

Definition 75 (Transition function sets & rate function)**.**

Let $\mathcal{F}^*_r = \{\mathcal{R}^*_{ij}\}, \mathcal{F}^*_c = \{\mathcal{C}^*_i\}$ *and* $\mathcal{F}^* = \{\mathcal{F}^*_r \cup \mathcal{F}^*_c\}$ *and define function* r^* *mapping* $\mathcal{F}^*$ *to* $\mathbb{R}$ *by, for any* $f \in \mathcal{F}^*$*,*

$$r^*(f) = \begin{cases} u_r & \text{if } f \in \mathcal{F}_r \\ u_c & \text{if } f \in \mathcal{F}_c. \end{cases} \tag{4.25}$$

Finally, we define an arbitrary bijection from $\mathcal{S}^*$ to $\{1, 2, ..., ||\mathcal{S}^*||\}$, and refer to both state s and its mapping as s, with the understanding that where it appears in an index we are referring to the integer.

We define the generator for our Markov chain to be matrix $\mathbf{Q}^* = [q^*_{sx}]$ where the non-zero off diagonals are given by

$$q^*_{sx} = \sum_{\{f \in \mathcal{F}^* : f(s) = x\}} r^*(f). \tag{4.26}$$

Just as in Section 4.1, the state space given above has a lot of redundancy in terms of biological interpretation of the model, and thus we reduce it in an anologous manner to the previous case.

Definition 76 (Permutation-equivalence)**.**

Let $\leftrightarrow$ be a binary relation on $\mathcal{B}$ defined by $\mathbf{A} \leftrightarrow \mathbf{B}$ if $\mathbf{A}$ can be transformed into $\mathbf{B}$ by any sequence of row and column swaps (permutations). For such $\mathbf{A}, \mathbf{B}$ we say that $\mathbf{A}$ and $\mathbf{B}$ are equivalent up to permutation.

In the case of $n \times n$ matrices, $\leftrightarrow$ can be conceptualised in terms of permutation matrices, where $A \leftrightarrow B$ if there exists a permutation matrix P such that $AP = B$.

Proposition 21.

$\leftrightarrow$ is an equivalence relation.

Proof.

Any matrix reached by some sequence of permutations can be returned to by the reverse sequence of permutations, thus $\leftrightarrow$ is symmetric. Further, any matrix relates to itself under the trivial permutation (not swapping any row or column), thus $\leftrightarrow$ is reflexive. Finally, transitivity can be seen by noting that if $\mathbf{B} \leftrightarrow \mathbf{A}$ and $\mathbf{A} \leftrightarrow \mathbf{C}$, then there exists a sequence of permutations mapping $\mathbf{B}$ to $\mathbf{A}$, and one mapping $\mathbf{A}$ to $\mathbf{C}$ — concatenating this sequence maps $\mathbf{B}$ to $\mathbf{C}$. □

Definition 77 (Binary Matrices up to permutations)**.**

We define $\mathcal{B}$ to be the quotient set of $\mathcal{B}^*$ by $\leftrightarrow$.

The binary matrices up to permutations have received some attention in the literature, e.g. Garriga et al. [44] focus particularly on finding representations with banded structure. However, for our purposes a banded structure offers no real advantage (since the matrices represent states, and are not used directly for computation tasks). The sequence of sizes of $n \times n$ binary matrices up to permutations is listed in [107]. There are 5624 such matrices for $n = 6$ (compared to $\approx 7 \times 10^{10}$ total binary matrices), and 251610 for $n = 7$ (compared to 6×10^{14}). Since we're primarily interested in $n \times z$ matrices for z and n of similarly small values, this is encouraging. Since our reduced state space (defined below) is a proper subset of the $n \times z$ binary matrices, this gives an upper bound on the number of states.

Further, if we considered just one of row or column permutations, the equivalence classes would be isometric to the symmetric group (see, e.g. [22]) S_{n*}, S_{z*} respectively, where n^*, z^* are the number of unique rows/columns. Since row and column swapping is clearly commutative, it follows that the equivalence classes of $\leftrightarrow$ are isomorphic to the permutation group $S_{n*} \times S_{z*}$, which is a subset of the symmetric group S_{z*n*}.

Thus, we can count the size of the equivalence classes via the orbit-stabiliser theorem, or Burnside's Lemma [22].

Both are informative results, but they do not offer much towards our model construction. To that end, we derive a few results ourselves;

Živković [135; 134] describes a procedure for finding a representative in terms of lexicographic orders on the rows and columns. We define a similar order on rows and columns to define a representative below.

Definition 78 (Representative matrix)**.**

For each class of $\mathcal{B}$, we define the representative matrix $\mathbf{A}$ as the member of its class with row order such that,

$$\sum_j \mathbf{A}_{ij} \leqslant \sum_j \mathbf{A}_{kj} \text{ for all } i < k, \tag{4.27}$$

and, for any $i < k$ for which the equality in (4.27) holds, then either the rows are identical, or

$$\min_j\{\mathbf{A}_{ij} = 1 : \mathbf{A}_{kj} = 0\} > \min_j\{\mathbf{A}_{kj} = 1 : \mathbf{A}_{ij} = 0\}, \tag{4.28}$$

and (similarly) column order such that

$$\sum_i \mathbf{A}_{ij} \leqslant \sum_i \mathbf{A}_{ik} \text{ for all } j < k, \tag{4.29}$$

and for and $j < k$ for which the equality in (4.29) holds, then either the columns are identical, or

$$\min_i\{\mathbf{A}_{ij} = 1 : \mathbf{A}_{ik} = 0\} > \min_i\{\mathbf{A}_{ik} = 1 : \mathbf{A}_{ij} = 0\}, \tag{4.30}$$

We say that $\mathbf{A}$ is the representative matrix of its class, or is in the representative form for its equivalence class.

That is, the representative matrix $\mathbf{A}$ is the member of its class for which the zero entries are as close to the northwest corner of the matrix as can be achieved via permutations.

Proposition 22.

The representative matrix of each class is unique.

Proof.

Let $\mathbf{A}$ be a representative matrix — that is, let $\mathbf{A}$ satisfy the four relations (4.27)–(4.29).

Consider swapping rows i and k ($i < k$) of $\mathbf{A}$ to form $\mathbf{C} \neq \mathbf{A}$ (of course, $\mathbf{C} \leftrightarrow \mathbf{A}$). Either $\sum_j \mathbf{A}_{ij} = \sum_j \mathbf{A}_{kj}$, or $\mathbf{C}$ does not satisfy (4.27).

If $\sum_j \mathbf{A}_{ij} = \sum_j \mathbf{A}_{kj}$, then $\mathbf{C}$ does not satisfy (4.28), since

$$\{\mathbf{A}_{ij} = 1 : \mathbf{A}_{kj} \neq 1\} = \{\mathbf{C}_{kj} = 1 : \mathbf{C}_{ij} \neq 1\},$$

and

$$\{\mathbf{A}_{kj} = 1 : \mathbf{A}_{ij} \neq 1\} = \{\mathbf{C}_{ij} = 1 : \mathbf{C}_{kj} \neq 1\}.$$

A similar argument applies for column swapping.

Thus any non-trivial row or column swap on $\mathbf{A}$ yields a matrix $\mathbf{C}$ which does not satisfy the relations (4.27)–(4.29).

Suppose now that the minimal sequence of swaps to map from $\mathbf{A}$ to $\mathbf{B}$ involves n_s row and z_s column swaps. Such a minimal sequence, in the sense that it involves the least number of total permutations, exists, and is unique up to the order of application of permuatations (as can be seen by considering the equivalence classes as permutation groups). If $n_s + z_s > 0$, then there is a row or column at which $\mathbf{A}$ and $\mathbf{B}$ disagree, and the same argument above shows that $\mathbf{B}$ does not satisfy the relations (4.27)–(4.29). Hence there is no $\mathbf{B} \neq \mathbf{A}$ for which $\mathbf{B} \leftrightarrow \mathbf{A}$ and $\mathbf{B}$ satisfies (4.27)–(4.29). □

Remark 10.

A simple algorithm for transforming a matrix $\mathbf{A}$ *into the representative form for its equivalence class can be achieved by performing iterations of the full* $n!$ *row-by-row and* $z!$ *column-by-column comparisons and swapping rows/columns according to Equations (4.27)–(4.29) (so that each iteration involves* $n! + z!$ *comparisons), stopping when an iteration does not instigate any swaps.*

For the small values of z and n we are interested in, the algorithm described in Remark 10 is efficient. With $n = z = 10$ the representative is evaluated in approximately 10^{-3} seconds on a Xeon X5650 desktop computer. We will not need to compute representatives for matrices much larger than 10×10, and for $n = z = 50$ the algorithm remains quite efficient, with computation times of approximately 10^{-2} seconds. An alternative is suggested by Živković [135], who proposed a branch and bound algorithm for finding representatives. However, for our purposes, the algorithm described in Remark 10 is sufficient.

We define the reduced model anaologously to the procedure in Section 4.1.2.

Definition 79 (Reduced state space $\mathcal{S}$)**.**

We define the state space for the reduced model by $\mathcal{S} = \mathcal{S}^* \cap \mathcal{B}$.

Next, we define the reduced transition functions corresponding to the biological processes of regulatory and coding null mutation, and rate function which associates these processes with appropriate contributions to the transition rate.

Definition 80 (Reduced transition functions)**.**

Let $\mathcal{R}_{ij}$ be a function with domain $\mathcal{S}$ defined by $\mathcal{R}_{ij}(s) = x$ such that $x \sim \mathcal{R}^*_{ij}(s)$ and $x \in \mathcal{S}$.

Also let $\mathcal{C}_{ij}$ be a function with domain $\mathcal{S}$ defined by $\mathcal{C}_{ij}(s) = x$ such that $x \sim \mathcal{C}^*_{ij}(s)$ and $x \in \mathcal{S}$.

Definition 81 (Reduced transition function sets & rate function)**.**

Let $\mathcal{F}_r = \{\mathcal{R}_{ij}\}, \mathcal{F}_c = \{\mathcal{C}_i\}$ and $\mathcal{F} = \{\mathcal{F}_r \cup \mathcal{F}_c\}$ and define function r mapping $\mathcal{F}$ to $\mathbb{R}$ by

$$r(f) = \begin{cases} u_r & \text{if } f \in \mathcal{F}_r \\ u_c & \text{if } f \in \mathcal{F}_c. \end{cases} \tag{4.31}$$

Finally, we define another arbitrary bijection mapping $\mathcal{S}$ onto $\{1, ..., ||\mathcal{S}||\}$ and refer to state s and its mapping under the bijection by s. Then the generator for our reduced state process is $\mathbf{Q} = [q_{sx}]$ where the non-zero off diagonals are given by

$$q_{sx} = \sum_{\{f \in \mathcal{F}: f(s) = x\}} r(f). \tag{4.32}$$

4.2.2 Efficiently computing the state space and generator matrix

We now consider the problem of computing the state space corresponding to n-genes and z-subfunctions, which we denote $\mathcal{S}^n_z$, and associated generator. To this end, a convenient notation for the states is given by specifying only the northwest corner s^* of state s where the other entries are all $1's$, i.e.

$$s = \left[\begin{array}{c|c} s^* & \mathbf{1} \\ \hline \mathbf{1} & \mathbf{1} \end{array}\right]. \tag{4.33}$$

The state corresponding to a matrix full of 1's (i.e. the state where all copies have a functioning version of all regulatory regions) is notated as [1].

A trivial, but computationally expensive method of generating the state space of the model is to iterate over the full set of binary matrices via the integers from 0 to $2^{nz} - 1$. Any matrices with one-or-more columns summing to 0 are rejected. We then use the algorithm described in Remark 10 to find the representative of each matrix's

equivalence class under $\leftrightarrow$, and add each previously-unencountered representative to a list. This is far from efficient, and iterating over the 2^{16} binary matrices for $z = n = 4$ took approximately 54 seconds on a Xeon X5650 desktop computer.

A more efficient approach computes the state space without directly referencing the binary integers. Starting from state [1] (from which all other states are accessible) we iterate over the possible transitions from each discovered state using the transition functions. We first apply the unreduced function and then put the resulting matrix into representative form. The list of states is updated as transitions leading to previously unidentified states are computed. This is advantageous since the generator can be formed at the same time using Equation (4.32). The time to compute the state space and generator for the case $z = n = 4$ (243 states), was 2.3 seconds, for $n = 6$ $z = 3$ (351 states) was 6.2 seconds, and for $z = 4$, $n = 5$ (948 states) was 33 seconds. However, this procedure can be further improved.

Consider, for example, the transitions from $[1] \to [0]$, which occur at rate nzu_r. The naive implementation of this approach references $\mathcal{R}_{ij}([0])$ for each i, j from 1 to n, z respectively. Each of the contributions of u_r from $r(R_{ij})$ is added to the generator after each reference. Each reference to $\mathcal{R}_{ij}([0])$ requires an application of the sorting algorithm to find the representative form [0], since $\mathcal{R}_{ij}([0])$ is itself defined by reference to $\mathcal{R}^*_{ij}([0])$. Thus this transition alone requires $2nz$ matrix additions (subtracting 1 from $[1_{ij}]$ and adding u_r to $[q_{[1][0]}]$) and nz applications of the sorting algorithm to transform the $n \times z$ matrix with a 0 in the ij entry into [0].

The $[1] \to [0]$ transition is the most expensive to compute, but the others also feature significant redundancy. Improvements can be made by counting the contribution to the rate using the fact that if row (and column) i (j) are identical to row (and column) k (l), then $\mathcal{R}_{ij}(s) = \mathcal{R}_{kl}(s)$ (Proposition 23 below). Hence only one such entry needs to be considered, with rates multiplied by the appropriate factor.

To that end, we define $\mathcal{E} = \{(i, j) : i = 1, ..., n, j = 1, ..., z\}$, and introduce an additional equivalence relation.

Definition 82 (Entrywise-equivalence)**.**

For fixed $\mathbf{A} \in \mathcal{B}^*$, we define binary relation $\sim$ on $\mathcal{E}$ by $(i, j) \sim (k, l)$ if swapping rows i and k of $\mathbf{A}$ yields $\mathbf{A}$, and swapping columns j and l of $\mathbf{A}$ also yields $\mathbf{A}$. We say that the (i, j) and (k, l) entries of $\mathbf{A}$ are equivalent.

Definition 83 (Entrywise-equivalent pairs)**.**

For a fixed $\mathbf{A} \in \mathcal{B}^*$ we define $\mathcal{E}^{\mathbf{A}}$ to be the quotient set of $\mathcal{E}$ by $\sim$. We choose a representative for each $E \in \mathcal{E}^{\mathbf{A}}$ by the member with least lexicographical order, i.e. by

$(i,j) \in E$ *where* $i = min_k\{(k,l) \in E\}$ *and* $j = min_l\{(i,l) \in E\}$.

Proposition 23.

If $(i,j) \sim (k,l)$ *then* $\mathcal{R}_{ij}(\mathbf{A}) = \mathcal{R}_{kl}(\mathbf{A})$ *and* $\mathcal{C}_i(\mathbf{A}) = \mathcal{C}_k(\mathbf{A})$.

Proof.

Since $(i,j) \sim (k,l)$, swapping row/column i,k / j,l of $\mathbf{A}$ yields $\mathbf{A}$. It follows that swapping rows i and k of $\mathcal{R}^*_{ij}(\mathbf{A})$ yields $\mathcal{R}^*_{kl}(\mathbf{A})$, and hence $\mathcal{R}^*_{il}(\mathbf{A}) \leftrightarrow \mathcal{R}^*_{kl}(\mathbf{A})$, thus $\mathcal{R}_{ij}(\mathbf{A}) = \mathcal{R}_{kl}(\mathbf{A})$.

A similar argument applies for operator $\mathcal{C}_i$. □

The northwest corner notation is convenient for counting entrywise equivalent pairs. The entries the in bottom right quadrant of Equation (4.33) can immediately be seen to be entrywise equivalent. The entries in the bottom left quadrant are equivalent if and only if their columns are equivalent, while those in the top right quadrant are equivalent if and only if their rows are equivalent, both of which can be tested using only the top left quadrant (northwest corner) s^*, which is (usually) a smaller matrix than s. It is easy to count the number of entries in each quadrant using only the size of s^*. If s^* is $n^* \times z^*$ then the bottom left quadrant is $(n - n^*) \times z^*$, the top right quadrant is $n^* \times (z - z^*)$, and the bottom right quadrant is $(n - n^*) \times (z - z^*)$.

Further, by storing the generator $\mathbf{Q}$ (and subgenerator $\mathbf{Q}^*$) in symbolic form, we can then refer to this whenever we update the parameters u_c, u_r to avoid computing the state space again. In MATLAB it is convenient to store as a complex matrix with $u_r = 1$ and $u_c = \sqrt{-1}$, which we denote $\mathbf{Q}^n_z$. In this case finding the generator for any previously considered combination of z, n is effectively instantaneous. We set $\mathbf{Q} = u_r \text{Real}(\mathbf{Q}^n_z) + u_c \text{Imag}(\mathbf{Q}^n_z)$. We have built a small library of these symbolic generator matrices (so far we have all combinations up to $z = n = 5$, with $\left|\left|\mathcal{S}^5_5\right|\right| = 5377$), which is updated whenever a previously unseen combination of z and n is computed.

With this, the procedure for computing with the model is then standard. However, it remains the case that computing the matrix exponential is extremely time-consuming. For $n = z = 5$ the computational time to calculate e^Q with $u_c = 3, u_r = 0.9$ was ≈ 10 minutes — clearly maximum likelihood estimation will require a high performance computer.

4.2.3 An alternative procedure for computing the state space

Another approach we considered was to directly compute the state space of $\mathcal{S}^z_{n+1}$ or $\mathcal{S}^{z+1}_n$ from the known state space of $\mathcal{S}^z_n$. With states identified as in Equation (4.33) it follows that $\mathcal{S}^{z_1}_{n_1} \subset \mathcal{S}^{z_2}_{n_2}$ whenever $n_1 \leqslant n_2$ and $z_1 \leqslant z_2$. We can consider the states gained by incrementing z or n, and use this to calculate the state space. We have so far only considered incrementing either z, or n, when the other is fixed to 2. This approach could further optimise the procedure discussed in Section 4.2.2. However, the procedure in Section 4.2.2 is sufficient to compute the appropriate state spaces, which are then stored for future reference, so we have set aside further development of this approach. Nonetheless, we outline our findings below.

Using the northwest corner notation, the state space for the case $z = n = 2$ can be written,

$$\mathcal{S}^2_2 = \left\{ [1], [0], \begin{bmatrix} 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \right\}, \tag{4.34}$$

with the states in the order they are written representing no functions having been lost, one function having been lost, two functions having been lost in the same copy, and two functions having been lost in separate copies respectively.

For $z = 3, n = 2$ we can write,

$$\mathcal{S}^2_3 = \mathcal{S}^2_2 \cup \left\{ \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \right\}, \tag{4.35}$$

where the additional states represent three functions being lost in one copy, and two functions being lost in one copy plus one in another copy.

For $n = 2$, $z > 2$, $\left|\left|\mathcal{S}^2_z - \mathcal{S}^2_{z-1}\right|\right| = \lfloor z/2 \rfloor + 1$. This can be seen by noting that all of the combinations of $z - 1$ or fewer 0's are already included in $\mathcal{S}^2_{z-1}$, so the additional states of $\mathcal{S}^2_z$ are all associated with z 0's. Correspondingly, the number of 0's each of the (unordered) rows could be associated with are: 0 and z; 1 and $z - 1$; 2 and $z - 2$; ... ; $\lfloor (z/2) \rfloor$ and $\lceil (z/2) \rceil$. Thus $\left|\left|\mathcal{S}^2_z - \mathcal{S}^2_{z-1}\right|\right| = \lfloor z/2 \rfloor + 1$. There is an obvious bijection between the unordered pairs and states introduced, with the larger of the pair corresponding to the number of columns with a 0 in the first row and a 1 in the second, and the smaller of the pair corresponding to the number of columns with a 1 in the first row and 0 in the second.

The pattern of incrementing n with $z = 2$ fixed is slightly less obvious, but after a few

enumerations it becomes quite clear, the first three cases are given by

$$\mathcal{S}_2^3 = \mathcal{S}_2^2 \cup \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \right\}, \tag{4.36}$$

and

$$\mathcal{S}_2^4 = \mathcal{S}_2^3 \cup \left\{ \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \right\}, \tag{4.37}$$

and the difference between $\mathcal{S}_2^5$ and $\mathcal{S}_2^4$ is

$$\begin{aligned} &\left\{ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \right\} \\ \bigcup &\left\{ \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \right\}. \end{aligned} \tag{4.38}$$

More generally, the difference between $\mathcal{S}_2^n$ and $\mathcal{S}_2^{n-1}$ is given by,

$$\left\{ \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}_{n-1}, \begin{bmatrix} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}_{n-1}, \begin{bmatrix} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 1 \end{bmatrix}_{n-1}, \ldots, \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}_{n-1}, \begin{bmatrix} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}_{n}, \ldots, \begin{bmatrix} 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 0 \end{bmatrix}_{n} \right\}, \tag{4.39}$$

where the subscripts denote the number of rows of s^* (as in Equation (4.33)), and for the last collection, the number of 1's in the first column never exceeds that in the

second. i.e. for $z = 10$,

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \tag{4.40}$$

is included but

$$\begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \tag{4.41}$$

is not. Thus, the number of states added by incrementing n with $z = 2$ is easy to calculate as

$$\left|\left|\mathcal{S}_2^n - \mathcal{S}_2^{n-1}\right|\right| = \begin{cases} 2 + (n-2) + (n-1) + (n-3) + (n-5) + \ldots + 1 & \text{for } n \text{ even} \\ 2 + (n-2) + (n-1) + (n-3) + (n-5) + \ldots + 0 & \text{for } n \text{ odd.} \end{cases} \tag{4.42}$$

In order to find an iterative procedure to define $\mathcal{S}_n^z$ for any n, z, we need to be able to either increment z for general n, or n for general z. The $n = 2$, $z = 2$ cases suggest that incrementing z for general n is likely to be easier, and this is supported by our numerical investigations up to now, which indicate that $\left|\left|\mathcal{S}_z^{n+1}\right|\right| > \left|\left|\mathcal{S}_{z+1}^n\right|\right|$.

Considering this, for any n incrementing z will add a set of states analogous to the case with $n = 2$ fixed, but instead of unordered pairs we have unordered n-tuples, still with fixed sum z. Counting these states is equivalent to counting the number of partitions of z into no more than n non-negative parts [23]. There is no explicit formula for the number of partitions, but generating functions are available. Regardless, it is convenient to consider the partitions one-by-one (and put them in bijection

with $\mathcal{S}^n_{z+1}$ at the same time). However, this is not the full collection of states added by incrementing z when $n > 2$. We ultimately set this analysis aside in favour of recursively computing the state space with the procedure described in Section 4.2.2.

4.2.4 Model for gene families of dynamic size

If we wish to model duplication events themselves, then the size of the gene family n is no longer fixed (the number of regulatory regions z still is, though).

Having already considered the case of a fixed-size gene family, extending the model to allow for additional duplication events is easy in principle. Here, we give a description of the necessary modifications to the model introduced in Section 4.2 to account for duplication.

We denote the state space for the case with n copies and z regulatory regions, and no additional duplication by $\mathcal{S}^n_z$ (see Equation (4.18) and Definition 79). Then, we define the state space for the case with additional duplication as follows.

Definition 84 (State space allowing for duplication)**.**

We define the state space $\mathcal{S}$ by

$$\mathcal{S} = \bigcup_{n=1}^{n_{max}} \mathcal{S}^n_z, \tag{4.43}$$

where n_{max} is the maximum allowed size of the gene family.

Transitions within states in $\mathcal{S}^n_z$ for fixed n occur with the same dynamics as in Section 4.2, except that we delete any row full of 0's.

We introduce an additional transition function to account for duplication.

Definition 85 (Duplication transition function)**.**

Define $\mathcal{D}_i(s)$ on $\mathcal{S}$ by the insertion of a copy of the i^{th} row of s into the $(i+1)^{th}$ position, with subsequent row positions incremented by 1. That is, $\mathcal{D}_i(s) = [\mathcal{D}_i(s)_{kj}]$, where

$$\mathcal{D}_i(s)_{kj} = \begin{cases} s_{kj} & \text{for } k \leqslant i \\ s_{(k-1)j} & \text{for } k = i+1 \\ s_{(k+1)j} & \text{for } k > i+1, \end{cases} \tag{4.44}$$

and $s = [s_{lj}]$.

We define $\mathcal{F}^{\mathcal{D}} = \{\mathcal{D}_i\}$, and we extend the rate function such that $r(f) = u_d$ for all $f \in \mathcal{F}^{\mathcal{D}}$, where we have assumed that duplication occurs at Poisson rate u_d for each

gene in the family. Then the generator is given by the same expression as before (Equation (4.32)), and a similar procedure to that which is described in Section 4.2.2 can be applied to compute it.

Remark 11.

In principle, $n_{max} = \infty$, however in practice, an appropriate truncation point can be found. The rate of loss to pseudogenization (and hence, transition from $\mathcal{S}_z^n$ to $\mathcal{S}_z^{n-1}$) will outweigh the rate of duplication (and hence, transition from $\mathcal{S}_z^n$ to $\mathcal{S}_z^{n+1}$) for any realistic parameters. Therefore, the process will enter the subset $\mathcal{S}_z^n$ with low probability when n is large, and hence truncation is appropriate.

4.3 Discussion

In this chapter, we have outlined some extensions of our model from Chapter 3 to account for a broader range of the potential dynamics of duplicate gene evolution. We will be conducting further analysis of the evolution of gene duplicates, and the models introduced in this chapter will be the starting points for this analysis.

In Section 4.1, we have outlined two models for the combined sub- and neofunctionalization process. Also, we have performed some initial analysis of the second sub- and neofunctionalization model. Our analysis indicates that any bias of our earlier parameter estimates (Section 3.8) from the presence of gene duplicates undergoing neofunctionalization would be minimal, even for a very high rate of neofunctionalization.

Further, we have fit this model to the *Homo Sapiens* and *Mus Musculus* data from Section 3.8. Based on this initial analysis, and subject to our modelling assumptions we find that the probability of preservation due to neofunctionalization is extremely low. We conclude therefore that neofunctionalization is not a major contributor to the preservation of gene duplicates, at least over the time scales over which regulatory subfunctionalization is resolved.

Neofunctionalization on the larger timescale would only be likely in the presence of some other factor leading to the preservation of both duplicates, and subfunctionalization is a likely candidate. Hence, our initial findings seem to support the hypothesis of Force et al. [42] that subfunctionalization acts to initially preserve duplicates giving neofunctionalization the opportunity to occur over larger timescales. Inclusion of a non-zero rate of neofunctionalization for nonfunctionalized regulatory regions, and analysis of data where neofunctionalization has been empirically identified would be illuminating.

In Section 4.2 we have outlined models for the subfunctionalization process to describe the evolution of gene families of fixed and dynamic size. We started by considering fixed size gene families in Section 4.2.1. We applied a procedure to reduce an intuitive model with extremely large state space to a model with much smaller state space. We then considered the computation of the model's state space, and its generator, and outlined an efficient procedure for doing so. We then extend the fixed-size family model to model duplication, and hence gene families of dynamic size, using the same procedure.

The model development procedure itself is somewhat more abstract than the standard approach, but circumvents the need to consider all of the possible transitions of the large state space in detail. We are ultimately able to define a complex model in terms of simple intuitions, together with some algebra.

It remains the case that numerical analysis of these models will be computationally expensive, since their state spaces, while vastly reduced from the intuitive models, is still quite large. In future, we will attempt to fit the dynamic-size gene family model to gene family data; clearly this will require the use of high performance computing infrastructure.

Chapter 5

Microsatellites

Microsatellites are an important class of genetic markers. They have been used to identify parentage or identity in forensic studies and to aid the construction of genomic maps [37; 21; 11]. They are also widely used in intraspecies studies to identify population structure and to infer demographic history [26; 105; 4; 111; 16].

In order to make accurate inferences about populations from microsatellite data, realistic models of how they evolve are required [37; 101]. The two most widely known models are the stepwise mutation model (SMM) of Ohta and Kimura [89] and the infinite alleles model (IAM) due to Kimura and Crow [64]. The SMM is generally thought to be a more realistic approach for microsatellite mutation, and is the underlying basis for most of the more detailed models. Models vary in complexity from simple random walks where the length (in number of repeat units) can change up or down by one, to more complex models where the probability of an expansion or contraction depends quadratically on the current length [21; 130].

As noted by Ellegren [37], the evolution of microsatellites is a complex process affected by many factors, these include the type of motif, the location in the genome, and the rate of point mutation. Point mutation can introduce impurities into the repeat sequence [47]. Furthermore, a variety of empirical studies have found evidence that impure repeat sequences undergo mutation due to slipped-strand mispairing at reduced rates relative to pure repeat sequences [74; 46; 125; 38; 57; 129].

Microsatellite data is regularly used in population studies, and many packages for demographic analysis include models for microsatellite mutation, for example BOTTLENECK [94], MSVAR [113; 8] and DIYABC [26]. While some software offers users the ability to choose between different microsatellite models, there are none which offer models that account for microsatellite purity explicitly.

Often in analysis, microsatellites are grouped together based on their motif-length. Cristescu [27] found that grouping pure and impure microsatellites separately was useful for inference on different timescales. In particular, Cristescu [27] noted that imperfect microsatellites may retain the signal for a bottleneck longer than more rapidly mutating perfect repeats.

Imperfect (or impure) microsatellites are thought to fit more closely to the infinite alleles model than pure microsatellites [25]. The infinite alleles model assumes that there are infinitely many possible different variations of a genetic sequence across a population. Intuitively, this is sensible, since there are a greater number of possible imperfect than perfect repeat sequences, and hence back mutation (which does not occur in the infinite alleles model) should be less likely. However, Harr [48] found that a slippage event in an impure microsatellite can result in the removal of an impure repeat. This suggests that not only can pure microsatellites become impure through point mutation, but impure ones can become pure once again through slippage, which indicates that pure and impure microsatellites should be treated together in a single model.

Some authors have attempted to reconcile the slipped-strand mispairing and point mutation processes, however these attempts have used schemes that disregard information about microsatellite purity in order to maintain a one-dimensional state space. Bell and Jurka [9] introduced the first model that attempted to include point mutation, they treated point mutation as a process that broke up a single microsatellite into two shorter ones. Kruglyak [68] treated point mutation as an event that truncates the microsatellite. Other authors have proposed variations on this scheme [106; 20]. Such models necessarily simplify the effect of the build-up of impurities through point mutation on the future evolution of microsatellites. These models do not account for the effect of purity on the rate of slipped-strand mispairing, and they cannot model the situation where a contraction in length removes an impure repeat unit, a situation which [47] found empirically does occur. This is particularly problematic since each of these models treats individual microsatellite as independent, but purifying mutations of the kind observed by Harr [47] effectively merge two pure microsatellites into one under these types of models, breaking the assumption of independence.

Cristescu [27] provides a convincing argument for the further investigation of imperfect microsatellites in the context of demographic studies. It is clear that the structure of microsatellites plays an important role in their evolution, and there is a divide between pure microsatellites, thought to fit best to derivatives of the stepwise-mutation model and impure microsatellites, which are usually handled with the infinite alleles model.

Here we will introduce a series of related models which we have developed to attempt to treat both pure and impure microsatellites by combining slipped-strand mispairing with point mutation. The model at which we ultimately arrive is then fit to whole-genome derived microsatellite data for nine animal species to test the hypothesis that interruptions in the repeat sequence lead to a decreased rate of slipped-strand mispairing, and to measure the size of this effect. The models themselves are essentially similar, with the distinction between them being mostly the consideration of how to handle fitting to the incomplete data which is available for microsatellite evolution. We start by introducing our initial model for microsatellite evolution, and then discuss the data-related considerations which led to further development of the model.

Remark 12.

As part of the research component of my honours degree [110], we investigated the existing microsatellite models extensively. We concluded by proposing a purity-dependent model which would extend the general model of Wu and Drummond [130] to account for interruptions in the repeat sequence. In keeping with the convention, this initial model was a CTMC, which would admit a stationary distribution for most parameters. Unconventionally, it was a level dependent quasi-birth-and-death process (LDQBD) [13], with the levels tracking the conventionally-modelled repeat number, and the phases tracking the extent of impurity in the repeat sequence. This chapter discusses our subsequent work refining and analysing this model.

5.1 Initial model

The model introduced in this section is a minor refinement of the one introduced in my honours thesis [110].

To model the evolution of an individual microsatellite with repeat motif of length L, we introduce the following process. Let $\{X(t)\}$ be a CTMC with state space given by

$$\mathcal{S} = \left\{(i,j) : i \in \{i_{\min}, i_{\min}+1, ..., i_{\max}\}, j \in \{0,1,...,j^i_{\max}\}\right\}, \tag{5.1}$$

where state (i,j) represents a microsatellite sequence comprised of i repeats of length L, j of which contain at least one mismatch to the motif.

There is no well defined rule to say which sequences are microsatellites [21]. The general consensus is that the minimum length threshold should depend on the motif length L, but not in such a way that $i_{\min}$ is fixed. As to how many impure repeat units should be allowed, there is very little discussion of this in the literature — $j = i$ is too large, since in this case every repeat unit is interrupted. If every repeat is interrupted,

the sequence is hardly repetitive, and would be unlikely to undergo slipped-strand mispairing at high rates. We leave aside the problem of defining cutoffs for now, but note that the minimum repeat number $i_{\min}$ is at least 2 and the maximum number of impure repeats $j^i_{\max} < i$. In principle the maximum repeat number $i_{\max} = \infty$, but in practice we will be truncating at some point, so we leave it unspecified.

$\{X(t)\}$ has generator matrix $\mathbf{Q} = [q_{(i,j)(k,l)}]$, where the non-zero off-diagonals are given by

$$q_{(i,j)(k,l)} = \begin{cases} r_s(i,j)\beta(i) & \text{for } k = i+1, l = j \\ r_s(i,j)(1-\beta(i))\frac{(i-j)}{i} & \text{for } k = i-1, l = j \\ r_s(i,j)(1-\beta(i))\frac{j}{i} & \text{for } k = i-1, l = j-1 \\ r_m(i,j) & \text{for } k = i, l = j+1, \end{cases}$$

the other off-diagonals all being zero. We will leave the expressions for the functions r_s, β, and r_m aside for now, but note that:

- $r_s(i,j)$ is the rate of slipped-strand mispairing, depending on the repeat number i and the level of impurity j. We assume that any slipped-strand mispairing event leads to a change in the repeat number by 1 (expansion) or -1 (contraction). We assume also that expansion events always lead to the introduction of a perfect copy of the repeat motif. We further assume that slippage events are not phase-altering, in the sense that they do not occur over multiple repeat units. That is, a sequence with motif length $L = 3$ could have the its first, second, and third nucleotides removed by a contraction, but not the second, third, and fourth, as this would alter the phase of the repeat unit. Out-of-phase nucleotides are treated as mismatches to the perfect sequence of repeats.

- $\beta(i)$ is the so-called bias function, with $\beta(i)$ giving the probability that, given a slipped-strand mispairing occurs, the event is an expansion (and $(1-\beta(i))$ the probability that the event is a contraction). β should be chosen such that $\beta(i) \to 0$ as $i \to \infty$, to reflect the empirically observed bias of long sequences towards contraction, and ensure numerical tractability.

- $r_m(i,j)$ is the rate at which impurities are introduced to the repeat sequence by point mutation. We assume that no point mutation events lead to the purification of a repeat sequence.

- The probability that a slipped-strand contraction event leads to the removal of an impure repeat unit is assumed to be equal to the proportion of impure repeat units j/i.

5.2 Microsatellite data

In principle, fitting a model such as defined above to some data should be very straight forward (once the various functions are specified). Given that the process has a stationary distribution, we can use this to calculate the likelihood of some parameters under the assumption that the observed distribution is at equilibrium.

Bright and Taylor [13] described a set of algorithms for calculating the stationary distribution (when such exists) of general level dependent quasi-birth-and-death processes (LDQBDs) based on the generalisation of Latouche and Ramaswami's logarithmic reduction algorithm. This method includes approximation for infinite state spaces, by truncating at some appropriately chosen level. Subsequently, Baumann and Sandmann [5] and Phuang-Duc et al. [93] proposed a memory efficient algorithm, based on a generalisation of continued fractions (with the two papers proposing essentially-equivalent algorithms). Besides these sophisticated approaches, the traditional methods (using e.g. Gaussian elimination) are adequate when the state space is not too large. We found that for this model, truncating at $i^*_{\max} = \max\{i : \beta(i) > \epsilon\}$ and solving via the MATLAB function mldivide (\) without exploiting the LDQBD structure gave consistent estimates to the combined algorithms of Bright and Taylor [13], and was computed in a shorter time (than my MATLAB implementation of the algorithms on my desktop computer).

However, a more fundamental problem exists for fitting this model to data, even when the state space is finite and the stationary distribution is easily calculated.

Implications of a systematic bias in microsatellite data

To define the state space explicitly (i.e. to choose $i_{\min}, j^i_{\max}$, and potentially $i_{\max}$) we need to decide which sequences should be regarded as microsatellites. There is no general agreement in the literature even insofar as how many repeat units are required to see characteristic microsatellite behaviour for uninterrupted sequences [21], let alone how many interruptions should be allowed before a sequence is no longer regarded as a microsatellite. It seems likely that there is no definitive answer to either of these questions, and that characteristic microsatellite behaviour would need to be identified on a per-locus basis, with any length or purity threshold used only as a guide in identifying such sequences.

We sought to sidestep this problem by noting that we would ultimately be working with a particular dataset provided by our colleague Dr. Bennet McComish, which he derived from publicly available whole-genome data (see Remark 13 and Table 5.1

below). The dataset comprises full descriptions of potential microsatellite loci for the whole-genomes of nine animal species. The genomes were searched using Tandem Repeats Finder (TRF) [10] with parameters chosen to be highly permissive of interruptions in the repeat sequence. By considering the algorithm used by TRF to identify repeat sequences, we define a boundary of precisely which sequences can be identified with our choice of TRF parameters, and use this to inform our choice of state space. There is no guarantee that TRF will identify any particular sequence, but in the absence of some error in the program's application of the Smith-Waterman algorithm, it should not report any sequences of a well defined kind (discussed below). Any microsatellite dataset derived from a whole-genome search by any particular software will have similar constraints, depending on the alignment algorithm used to score the sequences, and so this concept is generalisable to any whole-genome derived microsatellite dataset.

Remark 13.

The genomes Dr. McComish used to generate the TRF datasets were obtained from http://hgdownload.cse.ucsc.edu/downloads.html, aside from the penguin genome, which can be found at http://gigadb.org/dataset/100006.

The builds used for each genome are given in Table 5.1 below.

Common name	Scientific name	Build
lizard	*Anolis carolinensis*	anoCar2
chicken	*Gallus gallus*	galGal3
zebrafish	*Danio rerio*	danRer4
platypus	*Ornithorhynchus anatinus*	ornAna1
human	*Homo sapiens*	hg19
lancelet	*Branchiostoma floridae*	braFlo1
fruitfly	*Drosophila melanogaster*	dm3
nematode	*Caenorhabditis elegans*	ce10
penguin	*Pygoscelis adeliae*	PYGAD

Table 5.1: Genome builds used to generate the TRF datasets used in this analysis.

Tandem Repeats Finder (TRF) [10] is a program for searching DNA sequences of arbitrary length (including whole-genomes) for tandem repeat sequences, such as microsatellites. The TRF algorithm consists of two main components, a detection and an analysis component. The detection component uses a statistical model to detect sequences that are tandem repeats (with the possibility of interruption either by insertions and deletions or point mutations) with high probability. The analysis component,

and specifically its alignment procedure is of particular interest in terms of defining our model. For each of the detected tandem repeat sequences, a candidate pattern for the repeat motif is chosen from the sequence, and a Smith-Waterman style alignment [108] is computed between the candidate and each of the copies comprising the sequence. A second, so-called consensus pattern is chosen as the most frequently occurring series of matches during the initial alignment, and the alignment is performed again using the consensus pattern. Only those repeat sequences associated with alignment scores above some threshold in the second alignment are ultimately reported by the program, and thus any sequence for which the alignment to the consensus pattern scores below the threshold will not be reported.

The TRF parameters Dr. McComish used to generate the dataset were the following:

$$\begin{aligned} \text{match} &= 2, \\ \text{mismatch} &= -3, \\ \text{indel} &= -7, \\ \text{matching probability} &= 80\%, \\ \text{indel probability} &= 10\%, \\ \text{minimum alignment score} &= 18, \\ \text{maximum period} &= 6. \end{aligned} \tag{5.2}$$

The mismatch penalty of -3 is relatively low, as is the minimum alignment score of 18. This ensures that tandem repeat sequences with several mismatches to the pure repeat sequence will be included. The high indel penalty of -7 was chosen to attempt to minimise the number of sequences in the dataset with any insertion or deletion events (indels), which we do not model, and thus this dataset is highly appropriate for our purposes. Note that in many contexts, slipped-strand mispairing and various other events would be referred to as insertions or deletions — here we specifically mean insertions and deletions of a single neucleotide not corresponding to slipped-strand mispairing. When the best alignment is gap-free (i.e. the sequences do not differ by any insertions or deletions), then the alignment score is given by

$$S = \sum_{i=1}^{n} s(i), \tag{5.3}$$

where,

$$s(i) = \begin{cases} 2 & \text{if } a_i = b_i \\ -3 & \text{if } a_i \neq b_i. \end{cases} \tag{5.4}$$

Given that we do not model indels, we discarded all sequences which were associated with any insertions or deletion. Thus the alignment scores for the remaining sequences

should satisfy Equation (5.3), where A is the reported sequence and B is a sequence of the same length composed of perfect copies the consensus pattern reported by TRF (which we assume is the repeat motif). Since the number of mismatches in a sequence associated with state (i, j) in our model is between j and Lj, it follows that sequences associated with state (i, j) have alignment scores S such that

$$2iL - j(3 + 2L) \leqslant S \leqslant 2iL - 5j. \tag{5.5}$$

Therefore, any state for which

$$18 > 2iL - 5j, \tag{5.6}$$

would not be reported in the dataset, and we call such states *unobservable*.

On the other hand any state for which

$$18 \leqslant 2iL - j(3 + 2L), \tag{5.7}$$

would be guaranteed (assuming that it was detected during TRFs detection step) to be reported, and we call such states *observable*.

The rest of the states are associated with some sequences for which $S > 18$ and some for which $S < 18$ — we call these states *partially observable*. If we consider the TRF output as a sample of the state of sequences under our model, then the procedure is biased against partially observable states, with different levels of bias dependent on the state.

Ideally (from a fitting perspective), we would choose $i_{\min}$ and $j^i_{\max}$ such that all states in the model would correspond to observable sequences. However this would not be realistic from the modelling perspective, since we would expect mutations corresponding to transitions between observable and partially observable states to occur frequently. Since partially observed states will be under reported at various levels in the dataset, it may be difficult to fit this model to data. One possible approach would be to calculate the proportion of sequences which can be reported by TRF (with parameters given in Equation (5.2)) corresponding to each state, and calculate a modified distribution accordingly. On the other hand, the unobservable sequences are not particularly problematic. Mutations corresponding to transitions out of the space of unobservable states may be rare, in which case such states could be treated as absorbing. Alternatively, if such mutations are not rare, we can calculate the limiting distribution conditional on the process not being in an unobservable state. Thus, to proceed, it is desirable to redefine the model in such a way that all states are either observable, or unobservable — we will discuss this in Section 5.3 below. First, we discuss the procedure we performed to prepare the data for model fitting.

Post-TRF data handling

To summarise the procedure we used to prepare the data for model fitting, we

1. Grouped the data by genome and repeat length L — we have data for $L = 1, 2, 3, 4, 5, 6$.
2. Discarded all sequences with non-zero 'percent indels' as reported by TRF.
3. Performed pattern matching of the full sequence against the sequence of the same length composed of perfect copies of the 'consensus pattern' which was assumed to be the repeat motif, and determined the gap-free alignment score S.
4. Discarded the handful of sequences not obeying Equation (5.3).
5. Discarded any sequences with fewer than the minimum repeat-number thresholds of $6, 5, 4, 3, 3$ for sequences with motif-length $2, 3, 4, 5, 6$ respectively (the threshold for motif-length 1 sequences of 9 is imposed by Equation (5.3)).
6. Discarded the longest 1% of remaining sequences within each motif-length-genome group.

To justify this procedure, consider the following:

1. Motif length is thought to effect the dynamics of microsatellite evolution [37; 21].
2. An indel event would alter the phase of any downstream nucleotides, leading to significant miscounting of the number of mismatches. E.g. consider the sequence CATCATCAT, and suppose an insertion of neucleotide G occurs at the fourth position — the result is CATGCATCAT. The second sequence length is not an integer multiple of the repeat number, and without allowing for the possibility of gaps every character from the fourth position onward is a mismatch to the perfect repeat sequence.

   Compared to slipped-strand mispairing, indels are likely to be rare events, and accounting for them would require a more complex model again, with the model(s) under discussion here already representing a significant departure from the one-dimensional models in the literature. Thus, at least until we have a good model for point mutation and slippage alone, we will assume that no indels occur and fit to only gap-free sequence data.

   Note that for motif length $L = 1$, there is no way to distinguish between indels and mismatches, and the alignment algorithm with indel penalty -7 and mismatch penalty -3 will always choose the gap-free alignment for these sequences.

3. The 'consensus pattern' of length L reported by TRF is the best matching sequence to the majority of length L consecutive subsequences, and hence is the best candidate for the repeat motif.

4. Approximately one in one million of the remaining sequences failed to satisfy Equation (5.3). A likely explanation for this comes from the fact that we used TRF's report of 'percent indels' to identify sequences with indels. Rounding could lead to a large sequence with few (but non-zero) indels being attributed a 'percent indels' of 0.

5. These thresholds are in agreement with the base-pair counts of [128], rounded to the nearest integer. This is, in our estimation, fairly representative of the most commonly agreed on thresholds for characteristic microsatellite behaviour (e.g. see [19; 41; 71; 62]), although there is plenty of disagreement [21; 61]. Some authors have observed characteristic behaviour starting to occur in sequences with only one repeat unit and one extra nucleotide [73]. Allowing for such short sequences would result in a large proportion of the genome being identified as microsatellite sequences. We wished to choose a threshold that would minimise false identification of pure microsatellites, while allowing for at least one mismatches for most sequence lengths in the dataset; these thresholds achieved that.

6. The vast majority of sequences are distributed around the smallest repeat numbers (not many more than 10 for short motifs, and fewer for longer ones). However occasional observations of extremely-long (i.e. thousands of repeats) sequences also occur. We wanted a cutoff to remove the longest observations while preserving the majority of the data. In particular, we wanted to ensure that the data ultimately used for the fit contained at least some observations of all repeat-numbers included. We also expected long compute times for our fitting, and had plenty of data available, so reducing the size of the state space by excluding states corresponding to very-long and very-rare sequences was desirable. The shortest 99% of sequences achieved both of these criteria.

Most of these datasets (grouped by genome and repeat number) had empirical distributions which (by inspection) decreased approximately exponentially in repeat number. The distribution of mismatches within repeat number varied significantly, in some instances appearing normally distributed, exponentially increasing or decreasing, or roughly flat, and sometimes with spikes at some mismatch numbers. Figure 5.1 shows an example of the exponential-like decrease in repeat number. Some of the motif-length 2 and particularly motif-length 3 datasets had empirical distributions for which

repeat number (but not mismatches) was bimodal, an example of which can be seen in Figure 5.2. Many of the datasets had a large number of the shortest observable impure sequences — this is particularly pronounced in the motif-length 5 datasets, as well as some of the motif-length 4 and 6 had a large number short-impure sequences as seen in Figure 5.3. A similar phenomenon can be seen in Figure 5.1, but in this case the relative number of impure sequences is similar at larger repeat numbers.

Most of the datasets have more impure than pure repeats. In retrospect, this is indicative that our TRF parameters were likely too permissive of mismatches. We expect that the data is highly polluted with non-microsatellite sequence data (in the sense that many impure sequences were likely not descendants of sequences undergoing high rates of slipped-strand mispairing). However, this did not become clear until after the data fitting (discussed in Section 5.5).

5.3 An intermediate model

The approach we applied to deal with the variable bias against partially observable states was to redefine the model so that all states are either observable or unobservable. This is easily achieved by tracking the total number of mismatches instead of the number of interrupted repeats. The convention in microsatellite modelling is to track the number of repeats [37; 101; 123; 130], and so a natural extension of this would be to track the number of repeats which were impure. Such a model is in most senses simpler than one which tracks the number of mismatches, but when it comes to fitting to data, the discussion in Section 5.2 shows that the 'simpler' model is actually much more complicated — not to mention that it is also a lower resolution model of the biological process. To that end, we modify the model by interpreting j as the total number of mismatches to the perfect repeat sequence. The state space is defined by the same expression, but $j^i_{\max}$ could potentially be as large as iL (for the model in Section 5.1 it was no larger than i). More precisely, here $\{X(t)\}$ is a CTMC (specifically a LDQBD) with state space $\mathcal{S}$ given by

$$\mathcal{S} = \left\{(i,j) : i \in \{i_{\min}, i_{\min}+1, ..., i_{\max}\}, j \in \left\{0, 1, ..., j^i_{\max}\right\}\right\}. \tag{5.8}$$

State (i,j) corresponds to a sequence composed of i repeat units, having iL bases in total, j of which are mismatches to the corresponding sequence of perfect repeats. The process has generator matrix $\mathbf{Q} = [q_{(i,j)(k,l)}]$, where the non-zero off-diagonals are

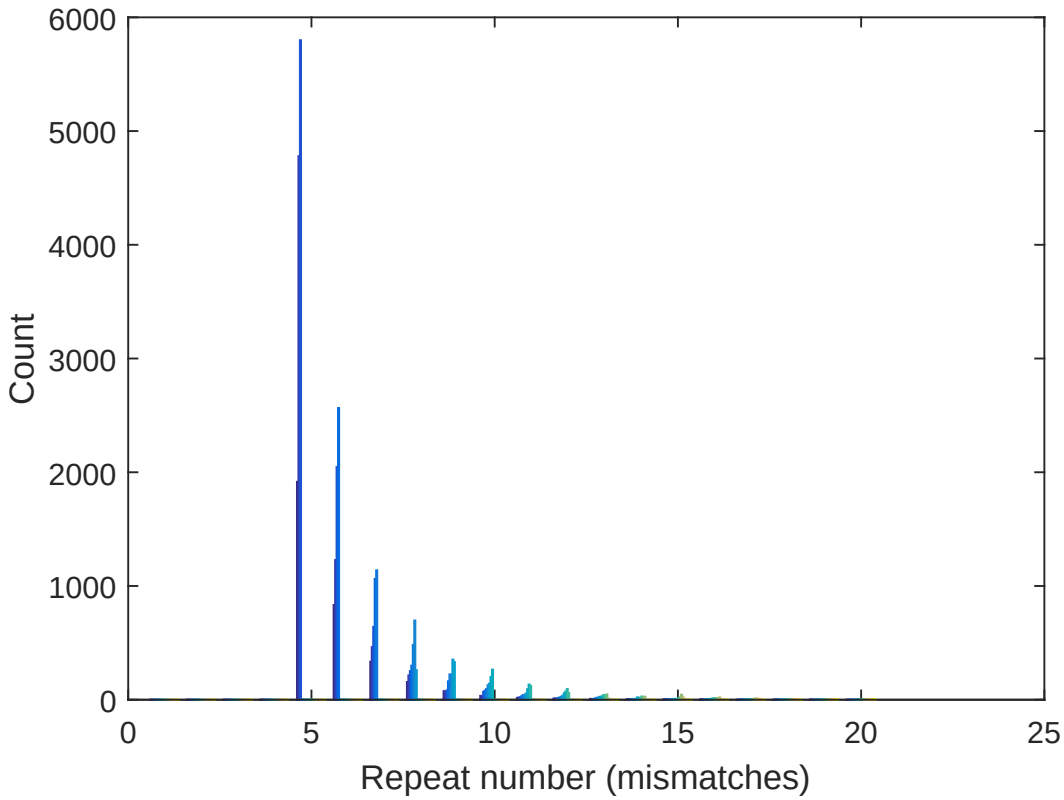

Figure 5.1: Histogram of counts of sequences with different numbers of mismatches to the pure repeat sequence, grouped by repeat number, for the motif-length 3 lancelet dataset. Each cluster of bars represents a single repeat number, with the leftmost (dark blue) bar representing the least-interrupted repeats, and the number of interruptions increasing to the right (light blue). This provides a typical example of the repeat-number distributions seen among our 54 datasets. The high numbers of impure sequences are likely due to misidentification of non-microsatellite sequences as interrupted repeats.

given by

$$q_{(i,j)(k,l)} = \begin{cases} r_s(i,j)\beta(i) & \text{for } k = i+1, l = j \\ r_s(i,j)(1-\beta(i))H(j-l, iL, j, L) & \text{for } k = i-1, j-L \leqslant l \leqslant j \\ r_m(i,j) & \text{for } k = i, l = j+1 \\ r_p(i,j) & \text{for } k = i, l = j-1. \end{cases} \tag{5.9}$$

In the first model (Section 5.1), a slipped-strand contraction event can potentially remove an impure repeat unit, and the probability that it does so (given such an event occurs) is modelled simply by the proportion of impure repeats j/i. This is equivalent to assuming that the location of the impure repeats and the location of the

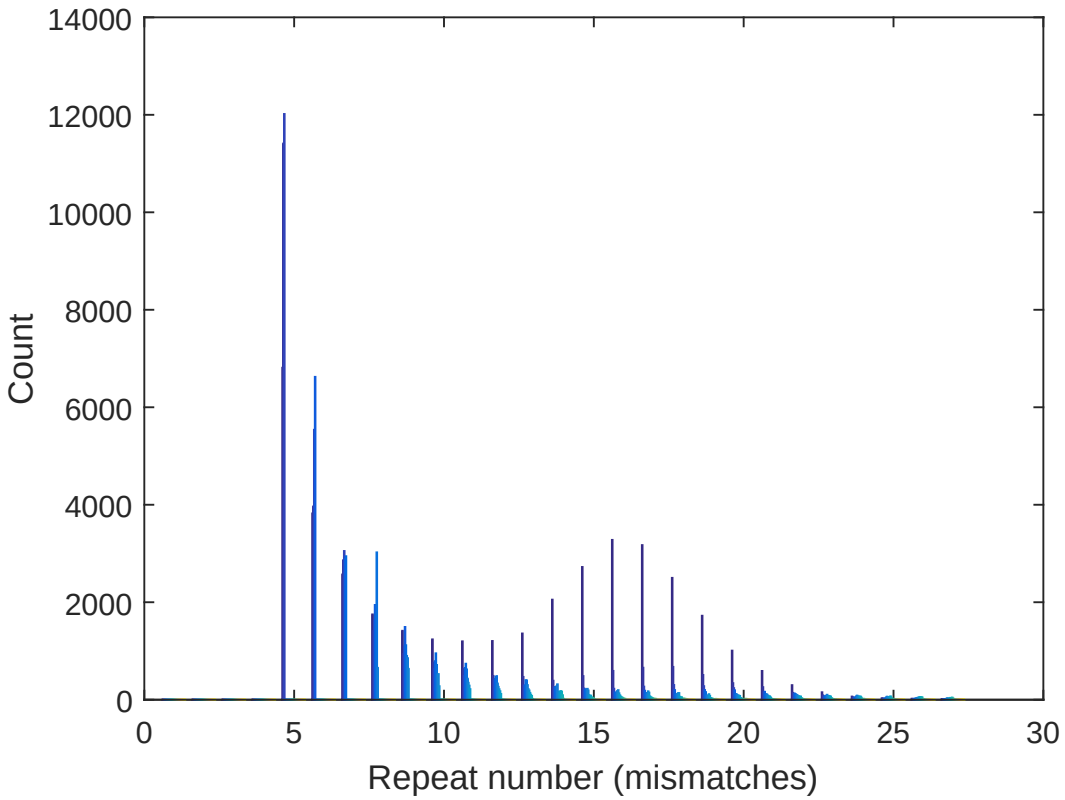

Figure 5.2: Histogram of counts of sequences with different numbers of mismatches to the pure repeat sequence, grouped by repeat number, for the motif-length 3 lizard dataset. Each cluster of bars represents a single repeat number, with the leftmost (dark blue) bar representing the least-interrupted repeats, and the number of interruptions increasing to the right (light blue). This shows an example of the bimodal-like repeat number distribution seen in some datasets.

repeat removed by a contraction event are both uniformly distributed.

Thinking of j as the number of mismatches now, it is possible for such an event to remove up to L mismatches. To account for this, we assume that the probability that such an event removes $j - l$ mismatches is given by the hypergeometric distribution $H(j - l, iL, j, L)$. The hypergeometric distribution $H(j - l, iL, j, L)$ describes the probability of $j - 1$ successes (mismatches) in L (number of bases removed) draws without replacement from a population of size iL (sequence length) containing exactly j successes (mismatches). As such, this is equivalent to assuming that the locations of the mismatches and the locations of the removed repeat unit are both uniformly distributed.

Since we track the exact number of mismatches, we now also account for the possibility

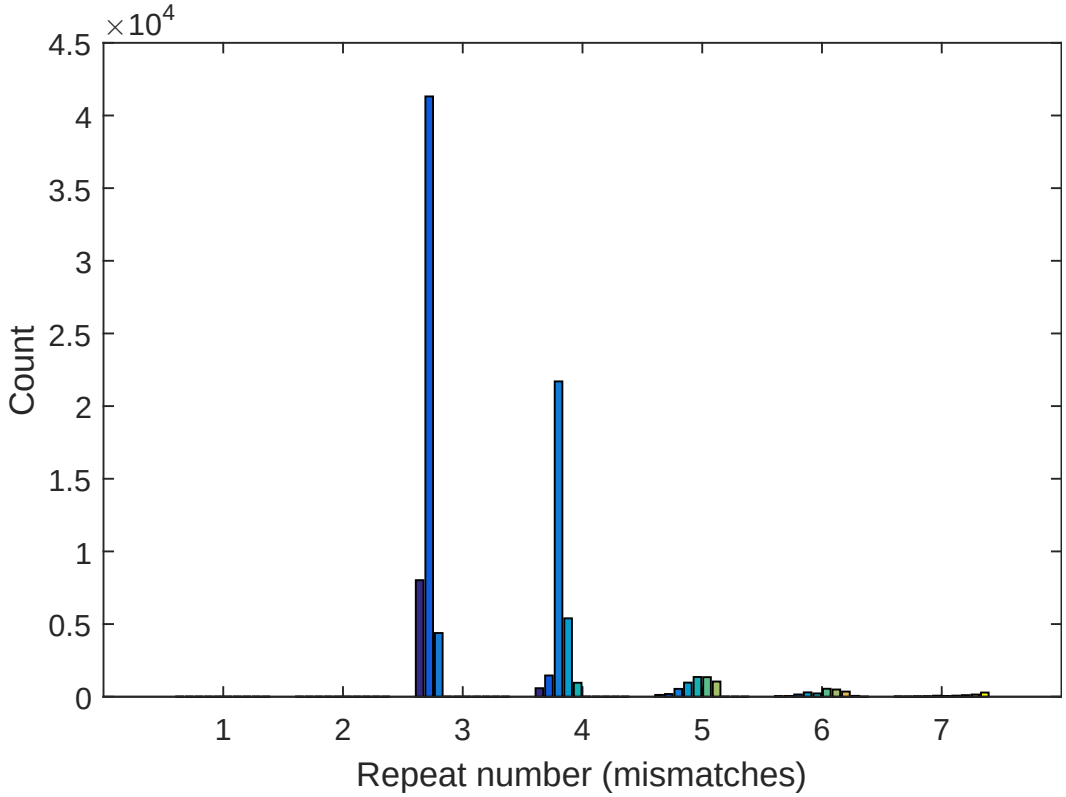

Figure 5.3: Histogram of counts of sequences with different numbers of mismatches to the pure repeat sequence, grouped by repeat number for the motif-length 5 lizard dataset. Each cluster of bars represents a single repeat number, with the leftmost (dark blue) bar representing the least-interrupted repeats, and the number of interruptions increasing to the right (light blue). This provides an example of the large number of short-impure sequences which are particularly prominent in the length 5 and 6 datasets.

that a point mutation event may remove a mismatch, by mutating the mismatching nucleotide back into a matching one. We did not include these events in the first model because we expect them to occur very infrequently (considering that two out of three of the potential nucleotide replacements are also mismatches). To purify an entire repeat unit could potentially take several such rare events, and we felt justified in excluding this possibility. With this model tracking individual mismatches there is no reason not to account for this, and so we assume that such mutations occur at a rate given by $r_p(i, j)$.

While the state space of this model is moderately larger than that of the first, it solves the problem of partially observable states in the dataset, and allows us to proceed with model fitting. We can ensure that all states are observable by choosing

$j^i_{\max} \leqslant \max\{j : 2iL - 5j \geqslant 18\}$ for all i, however it is not necessarily the case that we wish to exclude the unobservable states from the model entirely. To that end, we define the following subset of the state space.

Definition 86 (Set of unobservable states)**.**

We define $\mathcal{S}_u$ to be the set of unobservable states, that is

$$\mathcal{S}_u = \{(i, j) \in \mathcal{S} : 2iL - 5j < 18\}, \tag{5.10}$$

and we define $\mathcal{S}_0 = \mathcal{S} \backslash \mathcal{S}_u$.

When $||\mathcal{S}_u|| \neq 0$, we define the following limiting probability conditional on the process being in an observable state (conditional stationary distribution).

Definition 87 (Conditional stationary distribution)**.**

For each $j \in \mathcal{S}_o$, assuming the limit exists and is independent of i we define

$$\pi^o_j = \lim_{t\to\infty} \frac{1}{t} \int_0^t P(X(s) = j \mid X(0) = i, X(s) \in \mathcal{S}_o), \tag{5.11}$$

such that

$$\sum_j \pi^o_j = 1. \tag{5.12}$$

Given that the stationary distribution exists, the result

$$\pi^o_j = \frac{\pi_j}{\sum\limits_{i\in\mathcal{S}_o} \pi_i}, \tag{5.13}$$

is immediately obvious.

Restricting the state space to only include the observable states is unrealistic, since this would assume that the process of point mutation simply stops once a sequence is at the boundary of observability. Therefore, we conclude that we should include at least some unobservable states, and fit using Equation (5.13). However, this rationale can be extended to justify selecting $j^i_{\max} = iL$, since there is no reason for point mutation ever to stop. Any choice of cutoff for $j^i_{\max}$ will ultimately be arbitrary, and so we choose $j^i_{\max} = iL$ for exploratory analysis fitting this model to data.

Fitting the intermediate model to data was unsuccessful. The conditional stationary distribution is prone to numerical instability. Parameters can be chosen such that the stationary distribution has most of its weight in the unobserved states, leaving what essentially amounts to noise to be fit to the data via the conditional stationary distribution. Often, the resulting vector is not a probability distribution, since approximately

zero negative entries become moderately large. More often than not, these parameter sets (when they result in valid probability distributions) were selected as maximum likelihood estimates. Moreover, it is hard to trust the parameter estimates from such an unstable optimisation routine, even when the final estimate is well behaved. Besides these issues, it appeared that the likelihood was only locally optimal (we discuss this further in Section 5.5.1, where the final model had the same problem). The results of this fitting are mostly meaningless from a biological perspective. The preference for distributions with almost all of the weight in the unobservable states does, however, provide some suggestion that it is may be valid to treat the unobservable states as absorbing.

Moreover, we note that as long as $r(i,j)$ is decreasing rapidly in j, it follows that transitions leading from $\mathcal{S}_u$ to $\mathcal{S}_o$ are rare (but not the other way around). Therefore, we reasoned that the best approach was to treat the unobservable states as absorbing. The caveat is that it is not possible to use confidence intervals on the parameters of $r(i,j)$ to test the hypothesis that point mutation leads to a slowdown in the rate of slipped-strand mispairing (as we had originally intended). We effectively assume such effect by modelling the unobservable states as absorbing. Nonetheless, moving to an absorbing model was clearly the best way forward, and this approach is inline with the purported 'life-cycle' of microsatellites [85; 96; 114; 19]. This leads us to our final model.

5.4 Final model

The final model is a modification of the intermediate model, so that all of the states in $\mathcal{S}_u$ are considered absorbing (and collapsed into a single absorbing state A_m). We include two additional absorbing states, A_s and A_l corresponding to transitions which would lead to (i,j) with $i_{\min} > i$, and with $i_{\max} < i$ respectively. Absorption due to the repeat sequence becoming too short is a biologically realistic feature (e.g. Kelkar et al. [61] found that reduction in repeat number leads to 'death' of some microsatellites), while absorption due to the sequence becoming too long is included as a diagnostic to measure the success of truncation at $i_{\max}$. To summarise, in this case the model $\{X(t)\}$ is an absorbing LDQBD, with state space

$$\mathcal{S} = \big\{(i,j) : i \in \{i_{\min}, i_{\min}+1, ..., i_{\max}\}, j \in \mathbb{N}, j \leqslant \max\{j : 2iL-5j \geqslant 18\}\big\} \cup \mathcal{A}, \quad (5.14)$$

where $\mathcal{A} = \{A_s, A_m, A_l\}$, and transient state (i,j) represents a microsatellite consisting of i repeats, j of which are impure, and A_s, A_m and A_l are absorbing states representing a microsatellite below the minimum length threshold, above the (level dependent)

maximum mismatch threshold, or above the maximum length threshold respectively. The model has generator matrix $\mathbf{Q} = [q_{(i,j)(k,l)}]$, where the non-zero off-diagonals are given by

$$q_{(i,j)(k,l)} = \begin{cases} r_s(i,j)\beta(i) & \text{for } k = i+1, l = j \\ r_s(i,j)(1-\beta(i))H(j-l, iL, j, L) & \text{for } k = i-1, j-L \leqslant l \leqslant j \\ r_m(i,j) & \text{for } k = i, l = j+1 \\ r_p(i,j) & \text{for } k = i, l = j-1. \end{cases} \tag{5.15}$$

Here, $H(j-l, iL, j, L)$ is the hypergeometric distribution giving the probability of removing $j-l$ (which is between 0 and L) mismatches from a sequence of motif length L comprised of iL base pairs, j of which are mismatches. For notational convenience, we adopt the convention that transitions in Equation (5.15) which would otherwise lead to (i,j) with $i < i_{\min}$ or $i > i_{\max}$ instead lead to A_s and A_l respectively, and those which would lead to (i,j) with $j > \max\{j : 2iL - 5j \geqslant 18\}$ instead lead to A_m. We refer to the case in which $i_{\max} = \infty$ as the *untruncated process*. For the untruncated process, absorbing state A_l is inaccessible.

Now, we define subgenerator $\mathbf{Q}^*$ to be the block matrix of $\mathbf{Q}$ which records the transition rates between the non-absorbing states (i,j), and $\mathbf{V} = [\underline{\mathbf{v}}_{A_s}, \underline{\mathbf{v}}_{A_m}, \underline{\mathbf{v}}_{A_l}]$ to be a $* \times 3$ matrix with (column) vectors $\underline{\mathbf{v}}_{A_s}$, $\underline{\mathbf{v}}_{A_m}$ and $\underline{\mathbf{v}}_{A_l}$ recording the rates of transition to each of the absorbing states. We have, with $\mathbf{O}$ denoting zero matrices of appropriate sizes,

$$\mathbf{Q} = \left[\begin{array}{c|c} \mathbf{Q}^* & \mathbf{V} \\ \hline \mathbf{O} & \mathbf{O} \end{array}\right]. \tag{5.16}$$

5.4.1 Specifying the model

To specify the model we need to assume some particular expressions for $r_s(i,j)$, $\beta(i)$, $r_m(i,j)$ and $r_p(i,j)$. First, we assume that the rate of slippage, $r_s(i,j)$, depends linearly on the repeat number, and changes exponentially with the number of impurities, and so let

$$r_s(i,j) = (u_0 + u_1(i-1))c^j, \tag{5.17}$$

with the restriction that the parameters must take non-negative values. u_0 is a parameter for the base rate of slipped-strand mispairing, and u_1 is a parameter for the change in rate per repeat unit. c is a parameter for the reduction in slippage rate with impurity.

Linear dependence of slippage rate on repeat number has been used extensively in the literature, e.g. [101; 9; 68; 130]. Higher order functions have failed to provide better

fits than the simple linear dependence [101; 130], and this result was corroborated by our initial analysis of the existing microsatellite models (conducted as part of my honours degree [110]). A simple exponential decay function is conventional when accounting for rate-reducing effects, and hence we chose to account for impurity by multiplying our linear rate function by the factor c^j. Having a relatively simple (single parameter) dependence on purity keeps our function r_s manageable and avoids over-parameterization.

In the case that $c = 1$, assuming that the unobservable states are absorbing is equivalent to assuming that the buildup of impurities has no effect on the rate of slipped-strand mispairing until a threshhold value, at which point the process stops entirely. The threshold is assumed to be the at the boundary between the observable, and unobservable states. This is not likely to be a realistic model for microsatellite evolution, but it provides a point of comparison for the model with $c < 1$. We refer to the case with $c = 1$ fixed as the *purity-independent* model, with the understanding that we mean purity independent up to the cutoff of observably.

Next, we assume that the proportion of slippage events leading to expansions is given by a modified logistic function,

$$\beta(i) = \frac{1}{1 + e^{-(b_0+(i-1)b_1)}}. \tag{5.18}$$

Here, b_0 is the bias constant parameter, and b_1 is the bias linear parameter and; each may take any real value. The probability of contraction is given by $1 - \beta(i)$. This is the same function used by Wu and Drummond [130] to determine the proportion of expansion/contraction events. We found in our analysis of existing models (conducted as part of my honours degree [110]) that this bias function was the best performing when a suite of models was fit to whole-genome derived human, chimpanzee, and penguin data, as measured by the AIC.

To investigate the importance of the bias parameters, we consider two other variations of the model. The first is the *no-bias model*, in which $b_1 = b_0 = 0$, this fixes $\beta(i) = 0.5$ for all i, so that there is no bias. The second is the *constant bias model*, in which b_0 is free, but $b_1 = 0$, this results in an arbitrary constant bias. We will compare the no-bias and constant-bias models to each other, and to the the full model via the BIC (Equation (2.71)) in Section 5.2.

We further assume that the rate at which point mutations fix in each nucleotide is given by some constant $d \geqslant 0$, so

$$r_m(i, j) = d(iL - j). \tag{5.19}$$

Finally, we assume that a point mutation occurring in an already-mismatching nucleotide results in a change to a matching one with probability one third (since one of the three nucleotides it could mutate to is the correct one), hence

$$r_p(i,j) = \frac{1}{3}dj. \tag{5.20}$$

In the context where the relative content of nucleotides in the genomic region of interest is known $r_p(i,j)$ should be adjusted accordingly.

If d is fixed equal to 0 and $c = 1$ then the transitions between the levels are independent of the phase, and the phase is strictly decreasing, effectively reducing the model to a birth-death process. Thus, if we fix $j = 0$, $d = 0$, $c = 1$, the model reduces to a one-dimensional model with repeat number as its only dimension, and it is an absorbing (with A_s and A_l now the only accessible absorbing states in this case) version of the one-phase restriction of the general model introduced by Wu and Drummond [130]. If similar restrictions are placed on the intermediate model introduced in Section 5.3, then it is precisely the one-phase version of Wu and Drummond's model [130].

5.4.2 Limiting conditional distribution

In keeping with the standard procedure in microsatellite modelling, in which the stationary distribution of a CTMC is fit to the empirical distribution of allele lengths using likelihood based methods (e.g. [101; 130; 21]) our first idea was to fit the model to data using the Yaglom limit $\underline{y}$, or equivalently in this case, the unique quasi-stationary distribution (QSD) $\underline{\alpha}$ [31]. Recalling Definition 49, a distribution $\underline{\alpha}$ is called quasi-stationary if, for all $t \geqslant 0$ we have

$$\frac{\underline{p}_{\underline{\alpha}}(t)}{\underline{p}_{\underline{\alpha}}(t)\underline{1}} = \underline{\alpha}(t) = \underline{\alpha}. \tag{2.55}$$

For a finite irreducible state space, it is straight-forward to calculate the unique QSD using eigenvector decomposition of the subgenerator $\mathbf{Q}^*$. As demonstrated in [31], the QSD is the normalised left eigenvector of $\mathbf{Q}^*$ associated with the eigenvalue with largest absolute real part. This can be computed in various ways, often by using the MATLAB function eigs, which uses the Krylov–Schur Algorithm (due to Stewart [112]) to compute a subset of the eigenvectors. With the option 'LR' eigs returns only the eigenvector associated with the eigenvalue with the largest absolute real part. However, there are potential numerical difficulties in finding eigenvectors, although it is straight-forward to do so in principle.

Another approach is to use the return map, calculating the stationary distribution of the so-called 'returned process' in an iterative scheme. In the returned process,

transition to the absorbing state is followed by immediate return to the transient states according to some distribution $\underline{m}$. The returned process is a non-absorbing CTMC which, under suitable conditions, has unique stationary distribution, denoted $\underline{\pi}^{\underline{m}}$ [121]. Ferrari et al. [40] showed that when the process is finite and irreducible, the QSD $\underline{\alpha}$ is the unique solution to $\underline{\alpha} = \underline{\pi}^{\underline{\alpha}}$. Thus the QSD can be calculated by choosing initial distribution $\underline{m}_0$ and then setting $\underline{m}_n = \underline{\pi}^{\underline{m}_{n-1}}$.

Below, we describe an equivalent scheme, with the interpretation that the iterates are ratio of means distributions (Defintion 51). Specifically, α_{n+1_j} (defined in Equation (5.21) below) is the ratio of the expected time spent in transient state j to the expected time to absorption given initial distribution $\underline{\alpha}_n$. As noted by van Doorn and Pollett [121] same interpretation can be applied to the return map, so that if $\underline{\alpha}_0 = \underline{m}_0$, then $\underline{\alpha}_n = \underline{m}_n$ for all $n \in \mathbb{N}$. Van Doorn and Pollett [121] further note that this interpretation can be applied to calculate $\underline{\pi}^{\underline{m}}$. Here, we discuss this approach explicitly, and we show that under certain conditions the scheme (5.21) converges to $\underline{y}_{\underline{\alpha}_0}$, even when the transient set is reducible.

The iterative scheme can be performed as follows. Choose an initial distribution $\underline{\alpha}_0$, and for all $n = 1, 2, \ldots$ while $||\underline{\alpha}_{n+1} - \underline{\alpha}_n|| < \epsilon$, for some tolerance ϵ, let

$$\underline{\alpha}_{n+1} = \frac{\underline{\alpha}_n (Q^*)^{-1}}{\underline{\alpha}_n (Q^*)^{-1} \underline{1}}. \tag{5.21}$$

When the procedure stops, $\underline{\alpha}_n$ is the approximate QSD as desired.

The difference between the iteration of Equation (5.21) and the iteration of the return map $\underline{m}$ is how the iterates are calculated. When the return map is used explicitly, the stationary distribution of the returned process is calculated at each iteration. This is convenient as there are a range of efficient methods with which to find the stationary distribution. However, if $(\mathbf{Q}^*)^{-1}$ and $(\mathbf{Q}^*)^{-1}\underline{1}$ are computed once and stored, then each iteration of the scheme described above requires only three operations — being multiplication of a matrix by a vector, a vector by a vector and a vector by a scalar.

Next, we show that when the QSD is unique, it is the unique stationary point of the iterative scheme. A stronger result, guaranteeing convergence, is proved in terms of the return map by Ferrari et al. [40].

Lemma 4.

If absorbing CTMC $\{X(t)\}$ has unique QSD $\underline{\alpha}$, then it is the unique stationary point of the iterative scheme defined by Equation (5.21).

Proof.

First, we prove that the QSD is a stationary point of the iterative scheme.

Let QSD $\underline{\alpha}$ be the initial distribution of process $\{X(t)\}$. Then (as per [31]) we have, for any transient state i,

$$P(X(t) = i) = [\underline{\alpha} e^{\mathbf{Q}^* t}]_i. \tag{5.22}$$

Combining Equations (5.22) and (2.55) we have, for all $t \geqslant 0$,

$$\underline{\alpha} = \frac{\underline{\alpha} e^{\mathbf{Q}^* t}}{\underline{\alpha} e^{\mathbf{Q}^* t} \underline{1}} \tag{5.23}$$

$$= \frac{f(t)}{g(t)}. \tag{5.24}$$

However, $\underline{\alpha}$ does not depend on t, so can express $\underline{\alpha}$ exactly as a weighted average over any arbitrary values of t as

$$\underline{\alpha} = \frac{w(t_0)\frac{f(t_0)}{g(t_0)} + w(t_1)\frac{f(t_1)}{g(t_1)} + ... + w(t_n)\frac{f(t_n)}{g(t_n)}}{w(t_0) + w(t_1) + ... + w(t_n)}, \tag{5.25}$$

where the weight function $w(t)$ can be arbitrarily chosen (since $f(t)/g(t)$ does not depend on t, as per Equation (5.23)). In fact, we can write

$$\underline{\alpha} = \frac{\int_a^b w(t)\frac{f(t)}{g(t)} dt}{\int_a^b w(t) dt}. \tag{5.26}$$

Now letting $w(t) = g(t)$ for all $t \geqslant 0$ we have

$$\underline{\alpha} = \frac{\int_0^\infty f(t) dt}{\int_0^\infty g(t) dt} \tag{5.27}$$

$$= \frac{\int_0^\infty \underline{\alpha} e^{\mathbf{Q}^* t} dt}{\int_0^\infty \underline{\alpha} e^{\mathbf{Q}^* t} \underline{1} dt} \tag{5.28}$$

$$= \frac{\underline{\alpha} (\mathbf{Q}^*)^{-1}}{\underline{\alpha} (\mathbf{Q}^*)^{-1} \underline{1}}. \tag{5.29}$$

Thus, the QSD $\underline{\alpha}$ is a stationary point for the iterative scheme in Equation (5.21).

Next, we prove that any stationary point of the iterative scheme is a left eigenvector of $\mathbf{Q}^*$.

Suppose that $\underline{\alpha}^*$ is a stationary point, then

$$\underline{\alpha}^* = \frac{\underline{\alpha}^* (\mathbf{Q}^*)^{-1}}{\underline{\alpha}^* (\mathbf{Q}^*)^{-1} \underline{1}}, \tag{5.30}$$

letting $\underline{\alpha}^* (\mathbf{Q}^*)^{-1} \underline{1} = 1/\lambda$, we have

$$\underline{\alpha}^* = \lambda \underline{\alpha}^* (\mathbf{Q}^*)^{-1}$$

$$\implies \underline{\alpha}^* \mathbf{Q}^* = \lambda \underline{\alpha}^*. \tag{5.31}$$

Thus, $\underline{\alpha}^*$ is a left eigenvector of $\mathbf{Q}^*$, or at least a linear combination of eigenvectors corresponding to the same eigenvalue — from uniqueness of the QSD and the following argument we see that $\underline{\alpha}^*$ is indeed an eigenvector.

It follows from Theorem 1 of [121] that no eigenvector of $\mathbf{Q}^*$, besides the one associated with the unique QSD, can represent a probability distribution. Thus, if a stationary point of the iterative scheme is a probability distribution, then it is the QSD.

Finally, we prove by mathematical induction that if $\underline{\alpha}_0$ is a probability distribution up to multiplication by a constant (i.e. element-wise non-negative or non-positive, with at least one non-zero element), then $\underline{\alpha}_n$ is a probability distribution for all $n \in \mathbb{N}^+$.

Consider Equation (5.21) with $n = 1$

$$\underline{\alpha}_1 = \frac{\underline{\alpha}_0(\mathbf{Q}^*)^{-1}}{\underline{\alpha}_0(\mathbf{Q}^*)^{-1}\underline{1}}.$$

Clearly, $\sum_j \alpha_{1j} = 1$ as long as $\underline{\alpha}_n$ has a non-zero element.

Now we show that $\underline{\alpha}_1$ is element-wise non-negative, and hence is a probability distribution.

Suppose $\underline{\alpha}_0$ is element-wise non-negative. Then $\underline{\alpha}_0(\mathbf{Q}^*)^{-1}$ is element-wise non-positive, since $(\mathbf{Q}^*)^{-1} = -\int_{t=0}^{\infty} e^{\mathbf{Q}^* t} dt$ is element-wise non-positive. Further, $\underline{\alpha}_0(\mathbf{Q}^*)^{-1}\underline{1}$ is negative. It follows from Equation (5.21), that $\underline{\alpha}_1$ is element-wise non-negative.

The argument when $\underline{\alpha}_0$ is element-wise non-positive is analogous. In this case $\underline{\alpha}_0(\mathbf{Q}^*)^{-1}$ is element-wise non-negative, and $\underline{\alpha}_0(\mathbf{Q}^*)^{-1}\underline{1}$ is positive. Considering the form of Equation (5.21), we see that $\underline{\alpha}_1$ is element-wise non-negative.

Since $\underline{\alpha}_1$ is element-wise non-negative, and $\sum_j \alpha_{1j} = 1$, $\underline{\alpha}_1$ is a probability distribution. The same argument shows that when $\underline{\alpha}_n$ is a probability distribution, $\underline{\alpha}_{n+1}$ is also a probability distribution. Thus, by mathematical induction $\underline{\alpha}_n$ is a probability distribution for all $n \in \mathbb{N}^+$.

Therefore, the QSD is the unique stationary point of the iterative scheme.

□

In this thesis, we are interested in the case of an irreducible transient set, for which a unique QSD exists. Nevertheless, in the case of multiple QSDs, the proof can be extended. We show that, under certain conditions, if the scheme given by Equation (5.21) converges, it converges to $\underline{y}_{\underline{\alpha}_0}$.

Lemma 5.

Consider a regular absorbing CTMC $\{X(t)\}$ with initial distribution $\underline{\alpha}_0$, and Yaglom limit $\underline{y}_{\underline{\alpha}_0}$. For such a process, if the iterative scheme (5.21) (with initial distribution $\underline{\alpha}_0$), converges, then, given that

$$\lim_{t\to\infty}\lim_{n\to\infty}\underline{p}_{\underline{\alpha}_n}(t) = \lim_{n\to\infty}\lim_{t\to\infty}\underline{p}_{\underline{\alpha}_n}(t), \tag{5.32}$$

it converges to $\underline{y}_{\underline{\alpha}_0}$.

Proof.

The argument that any QSD is a stationary point of the iterative scheme is similar to the one given for the unique QSD in Lemma 4.

Theorem 1 of [121] shows that $\underline{\alpha}$ is a QSD if and only if it is x-invariant for $\mathbf{Q}^*$ for some $x > 0$. Equivalently, $\underline{\alpha}$ is a QSD if and only if it is a left eigenvector of $\mathbf{Q}^*$, associated with some real eigenvalue $\lambda < 0$, or it is a linear combination of such eigenvectors. Equation (5.31) shows that any stationary point $\underline{\alpha}^*$ is such. Thus, any stationary point of the iterative scheme is a QSD. Therefore, the set of fixed points of the iterative scheme is precisely the set of quasi-stationary distributions, with the condition that the starting vector $\underline{\alpha}_0$ is a probability distribution.

Next, we show that given the condition (5.32), the scheme (5.21) converges to the QSD associated with initial distribution $\underline{\alpha}_0$ (i.e. to $\underline{y}_{\underline{\alpha}_0}$). First, from Equation (5.21) with $n = 1$, we have

$$\underline{\alpha}_1 = \frac{\underline{\alpha}_0(\mathbf{Q}^*)^{-1}}{\underline{\alpha}_0(\mathbf{Q}^*)^{-1}\underline{1}}.$$

Now applying Proposition 15, we have

$$\begin{aligned}
\underline{p}_{\underline{\alpha}_1}(t) &= \frac{\frac{\underline{\alpha}_0(\mathbf{Q}^*)^{-1}}{\underline{\alpha}_0(\mathbf{Q}^*)^{-1}\underline{1}}e^{\mathbf{Q}^*t}}{\frac{\underline{\alpha}_0(\mathbf{Q}^*)^{-1}}{\underline{\alpha}_0(\mathbf{Q}^*)^{-1}\underline{1}}e^{\mathbf{Q}^*t}\underline{1}} \\
&= \frac{\underline{\alpha}_0(\mathbf{Q}^*)^{-1}e^{\mathbf{Q}^*t}}{\underline{\alpha}_0(\mathbf{Q}^*)^{-1}e^{\mathbf{Q}^*t}\underline{1}} \\
&= \frac{\underline{\alpha}_0 e^{\mathbf{Q}^*t}(\mathbf{Q}^*)^{-1}}{\underline{\alpha}_0 e^{\mathbf{Q}^*t}(\mathbf{Q}^*)^{-1}\underline{1}}.
\end{aligned} \tag{5.33}$$

Taking the limit as $t \to \infty$ we have

$$\begin{aligned}
\lim_{t\to\infty} \underline{p}_{\underline{\alpha}_1}(t) &= \lim_{t\to\infty} \frac{\underline{\alpha}_0 e^{\mathbf{Q}^* t}(\mathbf{Q}^*)^{-1}}{\underline{\alpha}_0 e^{\mathbf{Q}^* t(\mathbf{Q}^*)^{-1}}\underline{1}} \\
&= \frac{\lim_{t\to\infty} \underline{\alpha}_0 e^{\mathbf{Q}^* t}(\mathbf{Q}^*)^{-1}}{\lim_{t\to\infty} \underline{\alpha}_0 e^{\mathbf{Q}^* t}(\mathbf{Q}^*)^{-1}\underline{1}} \\
&= \frac{\underline{y}_{\underline{\alpha}_0}(\mathbf{Q}^*)^{-1}}{\underline{y}_{\underline{\alpha}_0}(\mathbf{Q}^*)^{-1}\underline{1}} \\
&= \underline{y}_{\underline{\alpha}_0}, \qquad (5.34)
\end{aligned}$$

where the final line follows from the fact that $\underline{y}_{\underline{\alpha}_0}$ is a fixed point of Equation (5.21). Therefore, from Proposition 16 we have $\underline{y}_{\underline{\alpha}_1} = \underline{y}_{\underline{\alpha}_0}$. The same argument can be applied for any $n \in \mathbb{N}$, so that $\underline{y}_{\underline{\alpha}_n} = \underline{y}_{\underline{\alpha}_0}$ for all $n \in \mathbb{N}$ by mathematical induction.

Now suppose that $\underline{\alpha}_n$ converges to some $\underline{\alpha}'$. That is, suppose

$$\begin{aligned}
\underline{\alpha}' &= \lim_{n\to\infty} \underline{\alpha}_n \\
&= \lim_{n\to\infty} \frac{\underline{\alpha}_{n-1}(\mathbf{Q}^*)^{-1}}{\underline{\alpha}_{n-1}(\mathbf{Q}^*)^{-1}\underline{1}}. \qquad (5.35)
\end{aligned}$$

Since $\underline{\alpha}'$ is a fixed point of Equation (5.21), it is a QSD, and hence $\underline{y}_{\underline{\alpha}'} = \underline{\alpha}'$. Applying Proposition 16 we have

$$\lim_{t\to\infty} \lim_{n\to\infty} \underline{p}_{\underline{\alpha}_n}(t) = \underline{\alpha}'. \qquad (5.36)$$

However, for any fixed $n \in \mathbb{N}$, we have $\lim_{t\to\infty} \underline{p}_{\underline{\alpha}_n}(t) = \underline{y}_{\underline{\alpha}_0}$, as per Equation (5.34). Thus, assuming Equation (5.32) holds, we have

$$\begin{aligned}
\underline{\alpha}' &= \lim_{t\to\infty} \lim_{n\to\infty} \underline{p}_{\underline{\alpha}_n}(t) \\
&= \lim_{n\to\infty} \lim_{t\to\infty} \underline{p}_{\underline{\alpha}_n}(t) \\
&= \lim_{n\to\infty} \underline{y}_{\underline{\alpha}_0} \\
&= \underline{y}_{\underline{\alpha}_0}. \qquad (5.37)
\end{aligned}$$

□

A sufficient condition for Equation (5.32) is that $\lim_{n\to\infty} \underline{p}_{\underline{\alpha}_n}(t)$ converges uniformly. In this case, the Moore–Osgood Theorem (see e.g. Lemma 5 of [84]) shows that the interchange of limits is valid. We conjecture that Lemma 5 could be strengthened — since $\underline{\alpha}_n$ is a weighted average of distributions, all of which are equal to $\underline{\alpha}_{n-1}\mathbf{P}(t)$ for some $t \geqslant 0$. This intuition suggests that $\left|\left|\underline{\alpha}_n - \underline{y}_{\underline{\alpha}_0}\right|\right| < \left|\left|\underline{\alpha}_{n-1} - \underline{y}_{\underline{\alpha}_0}\right|\right|$.

We now consider some numerical experimentation to compare the computation time of the scheme (5.21) to the equivalent scheme explicitly iterating the stationary distribution of the return map.

To test the convergence time of the scheme (5.21), we calculated the average time to find one QSD from a random starting vector using the iterative scheme with $\epsilon = 10^{-15}$, and the norm being the element-wise maximum absolute difference. For comparison, we also calculated the average time to compute the QSD with the standard implementation of the return map, i.e. using Gaussian elimination to calculate the stationary distribution of the returned process (with the same error), and using eigs (with default tolerance 10^{-14}). We averaged the compute times for

- 100 randomly generated processes with the property that the subgenerator matrix was 10×10 with n zero entries in the off-diagonals for each $n = 0, 1, \ldots, 90$;
- 10 randomly generated processes with the property that the subgenerator matrix was 100×100 with n zero entries in the off-diagonals for each $n = 0, 100, \ldots, 9000$; and
- 10 randomly generated processes with the property that the subgenerator matrix was 1000×1000 with n zero entries in the off-diagonals for each $n = 0, 10000, \ldots, 900000$.

The non-zero off-diagonal entries were all iid and drawn from a uniform distribution.

Between the ratio of means implementation (Equation (5.21)) and the return map (see e.g. [3]), the ratio of means implementation was universally associated with significantly reduced computation times. The ratio of means scheme (5.21) was also associated with reduced computation times compared to eigs for the 10×10 and 100×100 matrices, but with increased times for the 1000×1000 matrices. Also, the scheme (5.21) was prone to occasional very-slow convergence as compared to its median convergence time (observed mostly in the 10×10 case with more simulations). Table 5.2 shows the quartiles for compute time associated with each set of matrices and each numerical scheme.

We tested a few hundred randomly generated processes with 10000×10000 subgenerator matrices as well, and found that eigs was about 20 times faster than the iterative scheme in this case. We expect that the gap continues to widen as the matrices become larger. It should also be noted that for the case where multiple QSDs exist, the iterative scheme finds just the one QSD associated with its starting distribution. On the other hand, eigs can calculate the full set of linearly independent QSDs.

	Ratio of means			Return map		
	Q1	Median	Q3	Q1	Median	Q3
10x10	0.0004	0.0006	0.0025	0.0012	0.0019	0.0076
100x100	0.0006	0.0007	0.0009	0.0064	0.0071	0.0088
1000x1000	0.052	0.053	0.056	0.70	0.71	0.76

	eigs		
	Q1	Median	Q3
10x10	0.0023	0.0023	0.0023
100x100	0.0054	0.0064	0.0071
1000x1000	0.015	0.016	0.017

Table 5.2: Compute-time quartiles to calculate the QSD using iteration of ratio of means distribution, iteration of the the stationary distribution of the return map, and the MATLAB function eigs. Units are seconds, and computations were performed by a Xeon X5650 desktop computer.

Overall, the implementation (5.21) is much more efficient than explicit implementation of the return map. It is also competitive with the MATLAB function eigs for calculating a QSD when the number of transient states is smaller than 1000, and faster for small transient sets. The scheme (5.21) is convenient for finding the QSD associated with a particular intial distribution, since the natural implementation of the scheme does exactly that. If many QSDs are desired, eigs is clearly superior even for small transient sets, since the full set of linearly independent QSDs can be computed simultaneously. Besides this, the iterative scheme is stable whenever the inverse of the subgenerator can be calculated and the decay parameter is not very small relative to the precision of the computer.

A final note about the iterative scheme is that the m^{th} iterate is given by

$$\underline{\alpha}_m = \frac{\underline{\alpha}_0(\mathbf{Q}^*)^{-m}}{\underline{\alpha}_0(\mathbf{Q}^*)^{-m}\underline{1}}, \tag{5.38}$$

which is trivial to demonstrate by induction.

Thus the scheme can be modified to calculate only every m^{th} iterate, storing $(\mathbf{Q}^*)^{-m}$ and $(\mathbf{Q}^*)^{-m}\underline{1}$ and proceeding in similar fashion. Calculating powers of the inverse subgenerator matrix is time consuming, but powers of 2 are reasonably efficient to compute by iterating $(\mathbf{Q}^*)^{-2^{k+1}} = (\mathbf{Q}^*)^{-2^k}(\mathbf{Q}^*)^{-2^k}$ as discussed in Section 6.9 of Ross [100]. Another consideration is that the absolute value of $\underline{\alpha}_0(\mathbf{Q}^*)^{-n}\underline{1}$, and like-

wise the components of $\underline{\alpha}_0(\mathbf{Q}^*)^{-n}$, decreases rapidly with increasing m (see Remark 14 below). This leads to numerical instability if m is chosen too large (depending on the subgenerator matrix). Regular renormalising steps provide the necessary stability for practical computation.

We implemented the scheme with $m = 2, 4, 8, 16$ for simulations of the 10×10 and 100×100 matrices in the same fashion as above, except that we ran 100 simulations for each of the groups of 100×100 matrices. For the 10×10 matrices, computational time decreased with increasing m, however for the 100×100 matrices there was little difference. We tested a handful of 1000×1000 matrices and found that increasing m above 1 led to increases in computation time. Given that small matrices are already very fast to compute with, there seems to be little advantage to the scheme (5.38) compared to the scheme (5.21).

Remark 14.

For the case of a unique QSD, the decrease in $\underline{\alpha}_0(\mathbf{Q}^*)^{-n}$ is given by the decay parameter, being the negative of the eigenvalue $\mathbf{Q}^*$ with largest absolute value, as discussed in [121]. Unique or not, considering Equation (5.31) *we have*

$$\begin{aligned}
\underline{\alpha} =& \lambda \underline{\alpha}(\mathbf{Q}^*)^{-1} \\
=& \lambda(\lambda \underline{\alpha}(\mathbf{Q}^*)^{-1})(\mathbf{Q}^*)^{-1} \\
=& \lambda^2 \underline{\alpha}(\mathbf{Q}^*)^{-2} \\
=& \lambda^3 \underline{\alpha}(\mathbf{Q}^*)^{-3} \\
& \vdots \\
=& \lambda^n \underline{\alpha}(\mathbf{Q}^*)^{-n}, \qquad (5.39)
\end{aligned}$$

and λ is the decay parameter, or is an analogue to the decay parameter for the case where no communicating class associated with the eigenvalue of maximal real part is accessible from the starting distribution $\underline{\alpha}_0$.

It is also possible to exploit the LDQBD structure of the model to calculate the QSD. Discrete-time absorbing LDQBDs have recieved some attention in the literature [6; 7]. However, as shown by Kijima [63], and noted by Bean et al. [6], the QSD for the continuous time process is equal to the QSD of the embedded DTMC (Definition 28 of Chapter 2). Thus, the results in [6; 7] can be applied to calculate the QSD of the continuous time process by applying them to the embedded DTMC.

To truncate the infinite state space, we choose $i_{\max}$ equal to the maximum observed repeat number in the relevant subgroup of the dataset. Whenever the probability of

absorption into A_l is small, we reason that this represents a sufficient estimate of the QSD for the untruncated model (since the probability the process visits states above the level of truncation is low). If the probability of absorption into A_l is not sufficiently small, then we conclude that the parameters do not correspond to a good fit to the data anyway (since the untruncated process spends a non-negligible amount of time in biologically unlikely states), thus calculating the QSD with a higher truncation point is unnecessary. However, the QSD was not the distribution upon which we ultimately focused.

When the underlying process is not absorbing, it is justified to assume that given the relatively fast rate of evolution of microsatellites, and the long timescales over which evolution has been occurring they are highly likely to be observed at equilibrium frequencies in genomes [21]. On the other hand, if the underlying process is absorbing, there is no good reason to make this assumption. If microsatellites are born, evolve for a time under slipped-strand mispairing, and eventually die (as has been suggested, e.g. by Kelkar et al [61]), then the surviving microsatellites in contemporary genomes may be disproportionately likely to be relatively young, and it is not reasonable to assume that the population is at equilibrium. Thus, we define a different measure, being a transient distribution for a population of microsatellites each individually evolving by the process described above.

5.4.3 Extending the model to a population of microsatellites

So far, we have modelled the evolution of a single microsatellite over time. In practice, we will be working with data from a population of microsatellites observed at a single time, rather than a single microsatellite observed at many times. In the context of a non-absorbing stationary model, this is not a major concern — by assuming that the empirical distribution is (approximately) stationary, the procedure is simply to fit the stationary distribution of the model to the empirical distribution. Here we will make no such assumption, and we will fit the transient distribution of a population of microsatellites evolving under the model described in the previous section to the empirical distribution. In order to move away from stationary models, this consideration is paramount.

Remark 15.

Hautphenne et al. [49] have considered this problem (in a population biology context) in some detail and generality in their recent publication. Here, we have derived a similar result to Lemma 3.1 in [49] and applied it to calculate the probability of observing a microsatellite in a particular state at a particular time. The approach in [49] considers

more general birth processes, and this would be a great place to start from in the further development of non-stationary models both for microsatellites, and molecular evolution more broadly.

We first consider the birth process, which gives rise to individual microsatellites within the population. We assume a Poisson process to model the birth of microsatellites (i.e. we assume that births occur at a constant rate). Suppose that observation of the population of microsatellites occurs at a time t^* and let T_0 be a random variable tracking the time of birth of an individual microsatellite. Notice that a microsatellite of age t at time t^* must have been born at time $T_0 = t^* - t$. Since the birth process is Poisson, it follows that T_0 conditioned on $0 < T_0 \leqslant t^*$ is uniformly distributed [29; 100], and hence has density function,

$$\begin{aligned} f_{T_0}(t^* - t \mid T_0 < t^*) &= \lim_{\epsilon \to 0^+} \frac{P(T_0 \in I(t;\epsilon) \mid T_0 < t^*)}{\epsilon} \\ &= \frac{1}{t^*} \text{ for all } 0 < t < t, \end{aligned} \tag{5.40}$$

where $I(t;\epsilon) = (t^* - t - \epsilon, t^* - t]$.

If we observe a microsatellite at time t^*, we know not only that it was born before time t^*, but that it *survived* to time t^* — that is, it has not been absorbed before time t^*. Let T_a be a random variable tracking the time to absorption of an individual microsatellite. Then, ignoring any potential errors in the process of observation, the event that we observe a microsatellite at time t^* is the same as the event that the microsatellite was born before time t^*, and survived until at least time t^* — i.e. $T_0 < t^* < T_a$.

We wish to evaluate the probability that some microsatellite observed at time t^* is in some particular non-absorbing state s — this is the distribution we will ultimately fit to data. In order to evaluate this distribution, we will need to find the probability density function associated with the event that a microsatellite was born at time $T_0 = t^* - t$ given that it was observed at time t^*.

By the definition of the probability density function, we have

$$f_{T_0}(t^* - t \mid T_0 < t^* < T_a) = \lim_{\epsilon \to 0^+} \frac{P(T_0 \in I(t;\epsilon) \mid T_0 < t^* < T_a)}{\epsilon}. \tag{5.41}$$

Now, consider the probability that a particular microsatellite was born in the interval $I(t;\epsilon)$, conditional on that microsatellite having been observed at time t^*, given by

$$\begin{aligned} P(T_0 \in I(t;\epsilon) \mid T_0 \leqslant t^* < T_a) &= \frac{P(T_0 \in I(t;\epsilon), T_0 \leqslant t^* < T_a)}{P(T_0 \leqslant t^* < T_a)} \\ &= \frac{P(t_0 \in I(t;\epsilon), t^* < T_a, T_0 \leqslant t^*)}{P(T_0 \leqslant t^*, t^* < T_a)}. \end{aligned} \tag{5.42}$$

For notational convenience, we will denote $P(T_0 \in I(t;\epsilon) \mid T_0 \leqslant t^* < T_a)$ by $P(\ldots)$ within the context of this argument, and apply some equality preserving manipulation to the right hand side of Equation (5.42).

First we multiply by 1 to get,

$$P(\ldots) = \frac{P(T_0 \in I(t;\epsilon), t^* < T_a, T_0 \leqslant t^*)}{P(T_0 \leqslant t^*, t^* < T_a)} \frac{P(T_0 \in I(t;\epsilon), T_0 \leqslant t^*)}{P(T_0 \in I(t;\epsilon), T_0 \leqslant t^*)}. \tag{5.43}$$

Next, we apply the definition of conditional probability, and again multiply by 1, to get

$$P(\ldots) = \frac{P(t^* < T_a \mid T_0 \in I(t;\epsilon), T_0 < t^*) P(T_0 \in I(t;\epsilon), T_0 \leqslant t^*)}{P(T_0 \leqslant t^* < T_a)} \frac{P(T_0 \leqslant t^*)}{P(T_0 \leqslant t^*)}. \tag{5.44}$$

Again applying the definition of conditional probability (in both the numerator and denominator) we have

$$P(\ldots) = \frac{P(t^* < T_a \mid T_0 \in I(t;\epsilon)) P(T_0 \in I(t;\epsilon) \mid T_0 \leqslant t^*)}{P(t^* < T_a \mid T_0 \leqslant t^*)}, \tag{5.45}$$

applying an obvious corollary of Equation (5.40) we have

$$P(\ldots) = \frac{P(t^* < T_a \mid T_0 \in I(t;\epsilon)) \frac{\epsilon}{t^*}}{P(t^* < T_a \mid T_0 \leqslant t^*)}. \tag{5.46}$$

Now, we partion the region $(0, t^*]$ into n equal size regions of length ϵ_n, with the k^{th} such region denoted $I_k = ((k-1)\epsilon_n, k\epsilon_n]$ — note that the interval $I(t,\epsilon)$ is not one of the I_k's. We can think of $I(t,\epsilon)$ as a sliding region over the interval $(0, t^*]$ while the collection over k of I_k form a proper partition of the interval. However, we are free to choose $\epsilon = \epsilon_n$, and we do so. Now, we can apply the law of total probability to the denominator of Equation (5.46) to get,

$$\begin{aligned} P(T_0 \in I(t;\epsilon) \mid T_0 \leqslant t^* < T_a) &= \frac{P(t^* < T_a \mid T_0 \in I(t;\epsilon)) \frac{\epsilon}{t^*}}{\sum_{k=1}^{n} P(t^* < T_a \mid T_0 \in I_k) P(T_0 \in I_k \mid T_0 \leqslant t^*)} \\ &= \frac{P(t^* < T_a \mid T_0 \in I(t;\epsilon)) \frac{\epsilon}{t^*}}{\sum_{k=1}^{n} P(t^* < T_a \mid T_0 \in I_k) \frac{\epsilon}{t^*}} \\ &= \frac{P(t^* < T_a \mid T_0 \in I(t;\epsilon))}{\sum_{k=1}^{n} P(t^* < T_a \mid T_0 \in I_k)}. \end{aligned} \tag{5.47}$$

From Equation (5.41) we have,

$$\begin{aligned} f_{T_0}(t^* - t \mid T_0 < t^* < T_a) &= \lim_{\epsilon \to 0^+} \frac{P(T_0 \in I(t;\epsilon) \mid T_0 < t^* < T_a)}{\epsilon} \\ &= \lim_{\epsilon \to 0^+} \frac{P(t^* < T_a \mid T_0 \in I(t;\epsilon))}{\sum_{k=1}^{n} P(t^* < T_a \mid T_0 \in I_k)\epsilon} \\ &= \frac{\lim_{\epsilon \to 0^+} P(t^* < T_a \mid T_0 \in I(t;\epsilon))}{\lim_{\epsilon \to 0^+} \sum_{k=1}^{n} P(t^* < T_a \mid T_0 \in I_k)\epsilon} \\ &= \frac{P(t^* < T_a \mid T_0 = t^* - t)}{\int_0^{t^*} P(t^* < T_a \mid T_0 = t^* - u)du} \\ &= \frac{S(t)}{\int_0^{t^*} S(u)du}, \end{aligned} \tag{5.48}$$

where $S(t)$ is the survival function associated with the model described in Section 5.4. $S(t)$ gives the probability that the process has not been absorbed after evolving for a time t.

From Theorem 10, given in Chapter 2, we have,

$$S(t) = \underline{\alpha}_0 e^{\mathbf{Q}^* t} \underline{1}, \tag{5.49}$$

where $\underline{\alpha}_0$ is the initial distribution of process $\{X(t)\}$ for microsatellites born at any time t (which we assume does not depend on t), and $\underline{1}$ is a vector full of 1's. Intuitively, notice that $[\underline{\alpha}_0 e^{\mathbf{Q}^* t}]_s$ gives the probability that the process is in transient state $s = (i, j)$ at time t, and summing the probability the process is in any of the transient states gives the probability that it has not been absorbed. Thus, the probability density associated with the event that a microsatellite was born at time $T_0 = t^* - t$ given that it was observed at time t^* is given by

$$\begin{aligned} f_{T_0}(t^* - t \mid T_0 < t^* < T_a) &= \frac{S(t)}{\int_{t=0}^{T} S(t)dt} \\ &= \frac{\underline{\alpha}_0 e^{\mathbf{Q}^* t} \underline{1}}{\int_{t=0}^{T} \underline{\alpha}_0 e^{\mathbf{Q}^* t} \underline{1} dt} \\ &= \frac{\underline{\alpha}_0 e^{\mathbf{Q}^* t} \underline{1}}{\underline{\alpha}_0 (e^{\mathbf{Q}^* T} - \mathbf{I})(\mathbf{Q}^*)^{-1} \underline{1}}. \end{aligned} \tag{5.50}$$

Now, we wish to calculate the distribution, in terms of the Markov chain $\{X(t)\}$ defined in the previous section, of all of those microsatellites which survive to time t^*. Noting that a microsatellite born at time T_0 will have been evolving according to $X(t)$ for a time $t = t^* - T_0$ at the time of observation t^*. The probability that some

microsatellite is observed in a particular non-absorbing state s at time t^* is given by,

$$\begin{aligned}
&P(X(t^* - T_0) = s \mid T_0 < t^* < T_a) \\
&\quad = \int_{t=0}^{t^*} P(X(t) = s \mid T_0 = t^* - t < t^* < T_a) f_{T_0}(t^* - t \mid T_0 < t^* < T_a) dt \\
&\quad = \int_{t=0}^{t^*} P(X(t) = s \mid T_0 = t^* - t < t^* < T_a) \frac{\underline{\alpha}_0 e^{\mathbf{Q}^* t} \underline{1}}{\underline{\alpha}_0 (e^{\mathbf{Q}^* t^*} - \mathbf{I})(\mathbf{Q}^*)^{-1} \underline{1}} dt \\
&\quad = \int_{t=0}^{t^*} \left(\frac{[\underline{\alpha}_0 e^{\mathbf{Q}^* t}]_s}{\underline{\alpha}_0 e^{\mathbf{Q}^* t} \underline{1}} \right) \left(\frac{\underline{\alpha}_0 e^{\mathbf{Q}^* t} \underline{1}}{\underline{\alpha}_0 (e^{\mathbf{Q}^* t^*} - \mathbf{I})(\mathbf{Q}^*)^{-1} \underline{1}} \right) dt \\
&\quad = \int_{t=0}^{t^*} \frac{[\underline{\alpha}_0 e^{\mathbf{Q}^* t}]_s}{\underline{\alpha}_0 (e^{\mathbf{Q}^* t^*} - \mathbf{I})(\mathbf{Q}^*)^{-1} \underline{1}} dt \\
&\quad = \frac{[\underline{\alpha}_0 (e^{\mathbf{Q}^* t^*} - \mathbf{I})(\mathbf{Q}^*)^{-1}]_s}{\underline{\alpha}_0 (e^{\mathbf{Q}^* t^*} - \mathbf{I})(\mathbf{Q}^*)^{-1} \underline{1}}.
\end{aligned} \tag{5.51}$$

We can write this in vector form as

$$\underline{\pi}^*(t^*) = \frac{\underline{\alpha}_0 (e^{\mathbf{Q}^* t^*} - \mathbf{I})(\mathbf{Q}^*)^{-1}}{\underline{\alpha}_0 (e^{\mathbf{Q}^* t^*} - \mathbf{I})(\mathbf{Q}^*)^{-1} \underline{1}}, \tag{5.52}$$

where the s^{th} element of $\underline{\pi}^*(t^*)$ gives the probability of observing a microsatellite in state s at time t^*. Here we refer to both the state itself, and the integer to which it is associated with (by an implicit bijection from $\mathcal{S}$ to $\{1, .., ||\mathcal{S}||\}$) as s, with the understanding that the appropriate interpretation should be easily discernible from context.

Note that the limit as $t^* \to \infty$ of Equation (5.52) is given by

$$\lim_{t^* \to \infty} \underline{\pi}^*(t^*) = \frac{\underline{\alpha}_0 (\mathbf{Q}^*)^{-1}}{\underline{\alpha}_0 (\mathbf{Q}^*)^{-1} \underline{1}}. \tag{5.53}$$

Thus, if the time of observation is sufficiently large relative to the rate of mutation, the distribution approaches the ratio of means discussed in Section 5.4.2. This is in keeping with the physical interpretation of $\lim_{t \to \infty} \underline{\pi}^*(t^*)$ as the distribution of a population of microsatellites born with distribution $\underline{\alpha}_0$ which has been evolving for sufficiently long to reach equilibrium.

When $t^* = 0$, the expression in Equation (5.52) is undefined, however we note that the integral leading to Equation (5.52) is 0 when $t^* = 0$, so $\underline{\pi}^*(0) = \underline{0}$. This too is in keeping with the physical interpretation, since at time 0 the birth of microsatellites has not yet begun.

Non-dimensionalization of time

We can think of the parameters of our model as being of two distinct types — the parameters u_0, u_1 and d are rate-determining parameters, with u_0 and u_1 determining

the rate of slipped-strand mispairing, and d determining the rate of point mutation. The other parameters, b_0, b_1 and c are rate-modifying parameters, with b_0 and b_1 determining the ratio of slippage expansion/contraction events, and c determining the proportion of the relative rate at which sequences with different numbers of interruptions to the repeat structure evolve.

The rate-determining parameters warrant some further explanation, as it is important to realize that these are only meaningful relative to each other. Consider two sets parameters with distinct rate-determining parameters given by $\theta = [u_0, u_1, d]$ and $\theta' = [u_0', u_1', d']$ giving rise to generators $\mathbf{Q}$ and $\mathbf{Q}'$ respectively. Now suppose that $\theta = k\theta'$ for some $k \in \mathbb{R}$, and that each parameterization has the same rate-modifying parameters. It follows that $\mathbf{Q} = k\mathbf{Q}'$, since the rate-determining parameters are linear while the rate-modifying parameters act as coefficients to them. Notice from Equation (5.52) that

$$\frac{\underline{\alpha}_0(e^{k\mathbf{Q}^*t^*} - \mathbf{I})(k\mathbf{Q}^*)^{-1}}{\underline{\alpha}_0(e^{k\mathbf{Q}^*t^*} - \mathbf{I})(k\mathbf{Q}^*)^{-1}\underline{1}} = \frac{\underline{\alpha}_0(e^{\mathbf{Q}^*kt^*} - \mathbf{I})(\mathbf{Q}^*)^{-1}}{\underline{\alpha}_0(e^{\mathbf{Q}^*kt^*} - \mathbf{I})(\mathbf{Q}^*)^{-1}\underline{1}}. \tag{5.54}$$

Thus the constant t^* can be absorbed into the parameters u_0, u_1 and d through their scaling k (i.e. we can set $t^* = 1$ without loss of information). Hence Equation (5.52) becomes

$$\underline{\pi}^* = \frac{\underline{\alpha}_0(e^{\mathbf{Q}^*} - \mathbf{I})(\mathbf{Q}^*)^{-1}}{\underline{\alpha}_0(e^{\mathbf{Q}^*} - \mathbf{I})(\mathbf{Q}^*)^{-1}\underline{1}}. \tag{5.55}$$

Thus, we can fit the transient distribution directly to data without needing to treat time as a parameter of the distribution. On the other hand, if we were to fit an equilibrium distribution, we could by similar reasoning fix $d = 1$ and reduce the number of parameters of the model. As such fitting the transient distribution is still relatively computationally expensive than assuming that the process is at equilibrium. However, there is a lot to be gained by avoiding making the assumption that the empirical distribution is at equilibrium.

In particular, the non-dimensionalization of time provides a direct means of testing whether the empirical distribution really is at equilibrium. If the parameters inferred from fitting to data are such that $\underline{\pi}^* \approx \lim_{t^*\to\infty} \underline{\pi}^*(t^*)$ (given in Equations (5.53) and (5.55) respectively), then we can infer that the empirical distribution is at equilibrium. In this case, the relative clock d will not be informative, since any (larger) scaling k of the rate-determining parameters would lead to approximately the same fit.

On the other hand, if the inferred parameters are far from equilibrium, then we can make some inferences about the overall rate of mutation. The interpretation

of the Poisson rate parameter as an average number of events per unit time [100] is useful here. In this sense, provided that the process is far from equilibrium, the non-dimensionalization of time in Equation (5.55) is equivalent to scaling the rate-determining parameters such that d is the expected number of point mutations that the oldest microsatellite sequences (born close to time 0) in the genome would have undergone per nucleotide by the current time ($t^* = 1$). Thus, we can use d as a relative clock.

The interpretation for u_0, u_1 is philosophically similar, but less direct, since they are not precisely the rates of some mutational process, but rather phenomenologically-motivated parameters of rate $r(i, j)$, at which slipped-strand mispairing events occur. We can interpret $u_1 i + u_0$ as the average number of slipped-strand mispairing events the oldest microsatellites in the genome would have undergone per unit time during which they were in state $(i, 0)$, but this is not as helpful as the interpretation of d. Similarly, c^j can be interpreted as the proportion of slipped-strand mispairing events a sequence with j interruptions would undergo relative to an uninterrupted sequence of the same length.

The model at the population level combines the model for the evolution of individual sequences with a birth process. So far, we have thought of these as distinct, interacting models; this is convenient for formulation and analysis of the model. It is worth noting, however, that the combined process is itself a Markov chain. The state space of the population-level model is $\mathcal{S}^P = \{\mathbf{A} = [\mathbf{A}_{ij}] : [\mathbf{A}_{ij}] \in \mathbb{N}\}$, where $[\mathbf{A}_{ij}]$ gives the count of microsatellites in state (i, j). Represented as a Markov chain, the population level process is irreducible, and has stationary distribution equivalent to $\lim_{t^* \to \infty} \underline{\pi}^*(t^*)$ (given in Equation (5.53)), up to the total number of microsatellites (which is tracked by the population level process, but lost to $\lim_{t^* \to \infty} \underline{\pi}^*(t^*)$).

We assume that all birth events lead to a new microsatellite in the shortest and most pure state under the model $(i_{\min}, 0)$, such that $\underline{\alpha}_0 = \underline{e}_1$, a vector of $0's$ with a 1 in the position corresponding to state $(i_{\min}, 0)$ (which we define to be indexed at 1). Although this is unlikely to be strictly true in reality, the prevailing thought is that the majority of microsatellites begin their life-cycle as short and pure repeat sequences [18; 61]. Kelkar et al. found that deletion (in the broader sense, not the specific 1-nucleotide sense we use here) of interruptions was an important factor in the birth of longer microsatellite sequences [61]. Our model already accounts for events where interruptions are removed from a sequence of interrupted repeats, and such events are likely to account for many of the 'births' of longer microsatellites. With these considerations in mind, we think this assumption is a very reasonable one to

make. We note however that all that is required to account for some different model of microsatellite birth would be to adjust $\underline{\alpha}_0$ accordingly.

5.5 Fitting the model to genome data

We fit the distribution in Equation (5.55) to each of the datasets using a maximum likelihood approach and a combination of the MATLAB functions particleswarm and fminsearch (Nealder Mead Simplex). We did this for each of the submodels discussed in Section 5.4.1; the purity-dependent (c free) and purity-independent ($c = 1$ fixed) submodels, as well as the no bias ($b_0 = b_1 = 0$) and constant bias ($b_0 = 1$) submodels. We calculated the BIC (Section 2, Equation (2.71)) for each.

The combination of particle swarm and downhill simplex was used to search a wide range of the potential parameter space — we had no good intuition to direct the search, except that d should not be much larger than $u_0 + u_1$ and c should be between 0 and 1. Initial optimisation attempts suggested that downhill simplex alone was prone to becoming stuck in local optima, with the parameter estimates depending on the starting parameters chosen. Particle swarm (with local topologies) has the potential to search a wide range of parameters without becoming stuck in local minima [12]. Thus, we used particle swarm to first attempt to find the region of the global optimum, and then applied downhill simplex, to ensure that an optimum was found.

It should be noted that our search for parameter estimates encountered some significant identifiability issues (see Section 5.5.1), and the following results should thus be viewed with care.

The purity-dependent model achieved a better (lower) BIC than all others in 46 out of 54 cases. For the other 8 datasets the purity-independent model was preferred. These 8 were the motif-length 6 datasets for the human, penguin, fruitfly and zebrafish genomes, the motif-length 5 dataset for the lancelet genome, the motif-length 4 datasets for the nematode and fruitfly genomes, and the motif length 3 dataset for the nematode genome.

There were also 8 datasets for which the estimated value of c was greater than 1 (which would imply that increasing impurity increases, rather than decreases the rate of slipped-strand mispairing, and violate the assumption that the unobservable states can be treated as absorbing). Only one of these overlapped with the previous 8 discussed, being the motif-length 4 nematode dataset. The remaining 7 datasets for which the estimated value of c was greater than one were the motif-length 5 lizard, nematode, platypus and fruitfly datasets, the motif-length 6 lizard, chicken, and platy-

pus datasets. All 8 of these datasets had empirical distributions with a large amount of weight on short sequences with one or more mismatches, as in Figure 5.3. The model was not able to reproduce this behaviour. It is possible that many of the short, impure sequences in these datasets were misidentified as microsatellite sequences, and that the datasets were thus polluted with sequences which could not be expected to be well-modeled by the model presented here. Alternatively, our specification of the effect of mismatches in the repeat sequence might be insufficient for these cases.

We calculated the Kullback–Leibler divergence (Section 2, Equation (2.66)) from the theoretical distribution of the purity-dependent estimates to the associated empirical distributions as a measure of overall goodness-of-fit. We found that the KL-divergence for the datasets in which $c > 1$ was estimated, as well as for the datasets for which the purity-independent model was preferred by the BIC were significantly higher (mean 0.37 and 0.32 respectively) than the mean for the remaining datasets, which was 0.17. Hence it appears that the estimates for which the purity-independent model was preferred by BIC, and the ones for which a value of c greater than one was chosen were mostly cases of the model failing to fit the data reasonably well. Of the 30 datasets associated with a KL-divergence < 0.2, 3 had BIC scores favouring the purity-independent model, one of which was unique in having $c > 1$ ($c = 1.01$).

In total then, 39 of our 54 datasets yielded estimates consistent with a purity-dependent slowdown of slipped-strand mispairing. However, many were associated with relatively poor fits as shown by the KL-divergence. Of the 30 datasets associated with reasonably-good fits (KL-divergence < 0.2), 27 were consistent with purity-dependent slowdown of slipped-strand mispairing. All of the 30 datasets associated with KL-divergences < 0.2 were of motif-length 1–4, with $6, 7, 8$ and 9 (each out of 9) of the respective motif-length datasets associated with KL-divergences < 0.2. It should be kept in mind that this cannot be considered to be an unbiased test of the slowdown effect.

Eleven datasets were associated with estimates for which the rate of point mutation was larger than the rate of slipped-strand mispairing, at least for the shortest sequences — all but two of which overlapped with the purity-independent-preferred or $c > 1$ datasets discussed above. None of these were associated with KL-divergences less than 0.2, and the remaining datasets were all associated with estimates for which the rate of point mutation was either similar to, or an order of magnitude less than the rate of slipped-strand mispairing for the shortest sequences. Note that the rate of slipped-strand mispairing was assumed non-decreasing with sequence length, so the shortest sequences were associated with the lowest rates of slipped-strand mispairing. This

is consistent with the hypothesis that point mutation occurs at a rate one or several orders of magnitude lower than that of slipped-strand mispairing for microsatellite sequences.

Five datasets (all overlapping with the collection for which either $c > 1$ was estimated or the purity-independent model was preferred by BIC) had more than 1% probability of being absorbed into the A_l state used for truncation under the purity-dependent model. The truncation was chosen to match the state space of the model to the state space of the data, rather than being dynamically chosen based on parameters. Any significant absorption into A_l corresponds to this truncation having been inappropriate for the estimated parameters. Since we discarded the longest 1% of data we consider 1%-or-so absorption into state A_l to be acceptable, while results with more than 1% absorption into this state are likely to be associated with erroneous estimates of the transient distribution.

The values of c chosen for the majority of datasets were extremely small (median $5.54e^{-11}$) relative to the rate of slipped-strand mispairing (u_1, the linear slippage parameter had median 21). The parameter d associated with the rate of point mutation was, for the most part, close to 10 (15 of 30 datasets with KL-divergence < 0.2 and 21 of 54 total datasets had $9 < d < 11$). 2 of the 30 with low KL-divergence had d on the order of 10^3, while the rest were within an order of magnitude of 10. The parameter d can be interpreted as the average number of point mutations undergone by an average character in the sequence over the entire evolutionary history of the sequence, including reversals.

There did not appear to be any associations within the parameters of the model or between model parameters and motif-length. The KL-divergence scores for motif-lengths 5 and 6 were clearly much higher than for the other datasets, Figure 5.4 shows a box plot with KL-divergence against motif-length. However, we did not perform any tests for statistical significance of this difference; firstly because any such analysis would have been post-hoc, and secondly because the KL-divergence grouped by motif-length did not conform well to assumptions of common statistical tests.

Of the two models with restricted bias parameters, the constant-bias model was preferred by the BIC in 27 of 54 total cases, and 20 out of the 39 for which the full model had good KL-divergence scores. With the exception of 2 of the 27 cases in which the constant-bias model was preferred, a bias parameter equal to, or very close to 1 (corresponding to all slippage events being expansions) was chosen. These results were associated with higher rates overall rates of slipped-strand mispairing (u_1 had mean ratio 11 and ratio of means $2.7e^{-5}$) than the results for the full model, and with

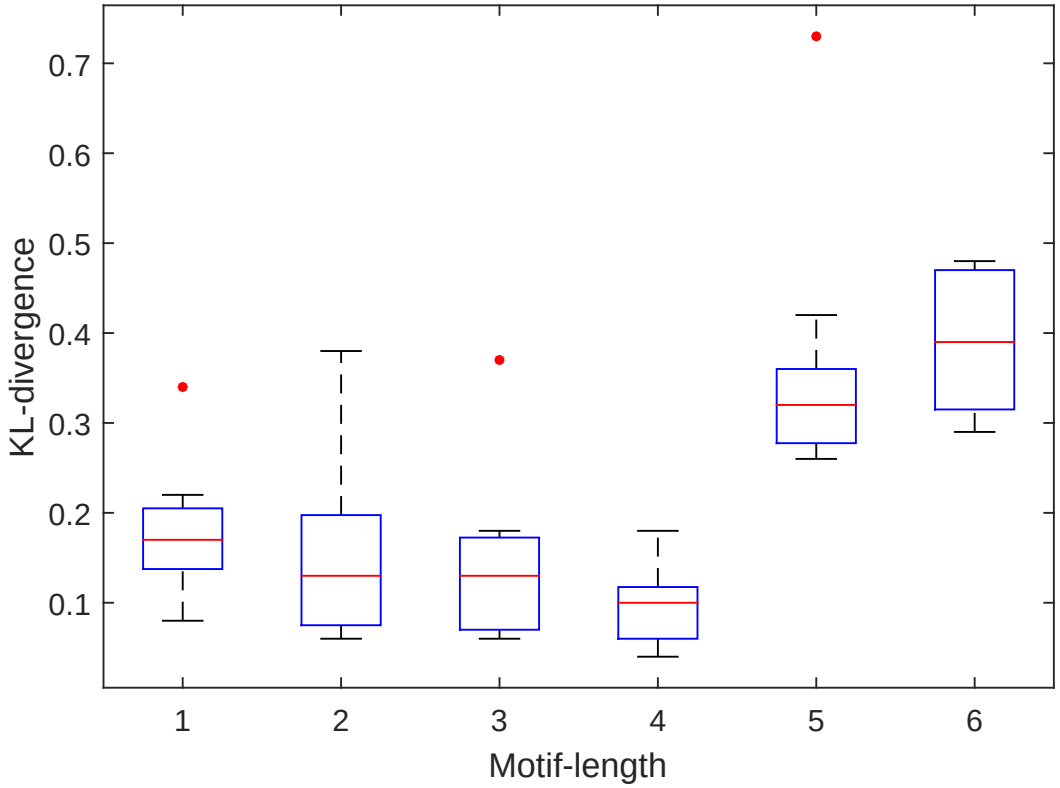

Figure 5.4: A boxplot showing the KL-divergence from the distribution of the model fit to the data, to the empirical distribution of the data.

substantially higher values of c (mean ratio $1e^{10}$, ratio of means 0.79), corresponding to a reduced effect of mismatch induced slowdown on the rate of slipped-strand mispairing — thus higher base rates of slippage, and much higher relative rates for impure sequences. They were also associated with higher rates of point mutation (d had mean ratio 5.69 and ratio of means $1.98e^{-7}$).

The no-bias model was preferred by the BIC in 10 out of 54 cases, 8 of which were associated with poor KL-divergence scores under the full model. The no-bias estimates were associated with high rates of absorption due to slipped-strand contraction events, which we expect are significantly overestimated under this model.

The full set of results is tabulated in Appendix A.

5.5.1 Model identifiability, and likelihood optimisation

The results of the fit to genome data were indicative of identifiability issues in the model parameter selection. Initial attempts to fit using the MATLAB function fmin-

search demonstrated that the (6 dimensional) likelihood surface had many local minima, with estimates being dependent on starting parameters. We used a combination of particle swarm and fminsearch to attempt to overcome this and find global minima, however results (Section 5.5) included some fairly unintuitive estimates. We performed a small suite of simulations with randomly generated parameters to test how reliably the combined particle swarm and downhill simplex could retrieve simulation parameters.

We found that the combined particle swarm and downhill simplex routine was insufficient to reliably estimate the simulation parameters, arriving at estimates that were substantially different to the simulation parameters in all ten simulations. The estimated distributions were close to the true distribution associated with the simulation parameters as measured by the KL-divergence (on the order of 0.01 for eight out of ten simulations). Thus, it appears that the model may not be identifiable. We did not observe any two parameter sets which gave rise to the same distribution exactly, and the model may be identifiable in the strict sense that the distribution is unequal when the parameters are unequal. However, the difference in the distributions can be very small, and in practice the model is not identifiable. This results in many local optima, and the optimisation routine fails to find the (possibly non-unique) global optima.

As such, the estimates inferred above should be viewed with some scepticism. Based on our simulations, even when the fit is good, the estimated parameters are likely to be only local, and not global optima. To find true maximum likelihood estimates (which are global optima) a careful analysis of the optimisation problem will be required.

With the parameters b_0 and b_1 fixed equal to the simulation parameters, the estimates for the remaining parameters were consistently close to the simulation parameters (at least up to scaling of the rate-determining parameters, as discussed in Section 5.4.3). This suggests that the logistic bias function is too flexible to be used in this model. It also suggests that the estimates associated with the no-bias version of the model are likely to be true maximum likelihood estimates, since model parameters are highly recoverable under this restriction. Even so, there are other reasons to be sceptical of these estimates, namely that the data appears to include many observations of non-microsatellite sequences, which is discussed further below.

5.6 Discussion

The results of this analysis are in-line with the predictions of a purity-dependent rate of slipped-strand in microsatellite sequences. 39 of 54 datasets were associated with

parameter estimates consistent with these predictions. However, the extremely-small values of c estimated suggest that the sequences with mismatches to the repeat motif are not undergoing slipped-strand misparing at any significant rate. Besides this, the estimates are not likely to represent true maximum likelihood estimates, since the optimisation routine did not appear to be finding global optima. The exception to this is the no-bias submodel, for which the optimisation routine was able to find global optima, and hence, maximum likelihood estimates.

Nonetheless, the estimates as they are suggest that a majority of the impure sequences in our datasets are not undergoing slipped-strand mispairing at any substantial rate. Almost all of the estimates had extremely low values of c, and this is true of the no-bias submodel just as it is for the others. Even for the submodels associated with low-quality estimates, there is no reason to think that the estimates should be biased towards low values of c, except insofar as they produce close fits to the data. The fact that almost every estimate was associated with very low values of c suggests that this part of the parameter space resulted in the best fits to these data.

This could be partly explained by the phenomena of old-microsatellite loci being killed off by the build up of point mutations. Intuitively we would expect these sequences to undergo a slow collapse, as slippage slows to the point of absence and the gradual process of point mutation eventually destroys the repetitive structure of the sequence — such a long death-cycle would lead to many such sequences being identified by TRF with our permissive parameter choice. It could also be partly explained by the misidentification of repetitive sequences which are not really impure microsatellites, in the sense of being a sequence which historically underwent high rates of slipped-strand mispairing and has picked up one or more mismatches to the motif through point mutation.

The very-low estimates for c provide some evidence that the absorbing model is a good choice for impure microsatellite sequences. Under these estimates, the mutation rate is effectively 0 well before the level-dependent absorbing boundary for all levels, and as such assuming an absorbing boundary is well justified. The caveat is that this could come down to the pollution of the data with non-microsatellite sequences. Suppose that genuine microsatellite sequences in state (i, j) continue to evolve with characteristic microsatellite behaviour. In this case, the boundary for absorption at level i should occur at some $j' > j$. However, it remains a possibility that the majority of sequences detected in state (i, j) would not be well classified as microsatellite sequences, and would not undergo slipped-strand mispairing at any significant rate. If this is the case, then our estimation of the rate of slipped-strand mispairing associated

with state (i, j) will be significantly biased towards 0. We would, under such circumstances, simultaneously underestimate the appropriate absorbing boundary j', and overestimate the corresponding parameters of the microsatellite detection software.

Either way, we suspect that the effect of impure sequences in our data which are undergoing high rates of slippage has been dominated by sequences which are not. This is supported by the estimates of parameter c, as well as by the observation that a majority of sequences in the dataset contain at least 1 interruption. Between this, and the quality of the estimates of most of the submodels, we do not expect that our estimates are representative of the behaviour of genuine microsatellite loci. A further consideration based on this result is that the rate of slowdown by slipped-strand mispairing might be better modelled with a steeper rate of decline than exponential, for example, by c^{j^a}. In this way, very low rates of slipped-strand mispairing can be attributed to highly impure sequences while slightly impure ones can retain a fairly high rate.

It is likely that stricter alignment scoring parameters are required for whole-genome derived sequences than we used here in order to reduce the probability of non-microsatellite sequences being reported. Further investigation with empirically-confirmed microsatellite sequences, rather than the whole-genome derived sequences used here would be very valuable. Fitting the model (with a more reliable optimisation procedure) to data from pure and impure sequences which are confirmed to be evolving with characteristic microsatellite behaviour would yield estimates more representative of bona fide impure microsatellites.

In some instances, the model failed to provide a good fit to the data as measured by the KL-divergence from the theoretical to the empirical distribution. 30 of the 54 datasets resulted in KL-divergences less than 0.2, which represents a fairly good fit, on the other hand some had scores as high as 0.7, which represents a very poor fit. The worst fits were associated with the datasets for which parameter estimates were not consistent with the predictions of a purity-dependent rate of slipped-strand mispairing.

The scores for the motif-length 5 and 6 datasets appeared to be substantially worse than for the shorter datasets. We suspect this was caused by our choice of very-permissive TRF parameters resulting in samples polluted with sequences which would not be well characterised as microsatellites. Intuitively it would make sense for this to be a bigger problem for longer-motif repeat sequences, since a sequence of a fixed length is more repetitive when the motif is shorter. We did not investigate this further, as doing so would be little more than data-dredging at this point in the analysis.

We hypothesise that a steeper motif-length dependent cut-off is appropriate for identifying impure microsatellite sequences than the sequence-length cut-offs suggested by previous work on pure sequences, and this could be tested (on a different dataset, since this data was used to generate the hypothesis) by fitting a purity-dependent model to data with varying cut-offs and comparing the fit.

We found that the logistic bias function did not perform much better than a constant bias as measured by the BIC. The constant bias was almost always selected so that every slippage event was an expansion. Moreover, the bias function appeared to be largely responsible for the identifiability issues we experienced in fitting the model to data. It seems likely that the interaction between purity-based slowdown of the slippage rate, and absorption by accumulation of impurities provides a sufficient alternative to a bias towards contractions in terms of the distributions which are attainable under the model.

An intuitive explanation for this comes from considering a realisation of the individual-level process under the assumption that every slippage event is an expansion, and $c \approx 0$. In this case, the process is essentially a randomly-killed pure birth process. Initially, the process is in state $(i_{\min}, 0)$, and has a low rate of transition to $(i_{\min}+1, 0)$, or to $(i_{\min}, 1)$. Until the first phase-transition, the process behaves like a pure-birth process, and transitions in the level variable occur at a monotonically increasing rate. Thus, the expected time spent in each level is less than the last. As soon as the first phase-transition occurs it is very unlikely that any further level-transitions will occur before absorption, since $c \approx 0$ ensures that such transitions occur at approximately zero rate. Some flexibility in terms of the relative time spent in each level is afforded through the rate-determining parameters u_0, u_1 and d. Thus, although all of the slipped-strand mispairing events are expansions, the inclusion of the phase variable allows for the same kinds of distributions that would be attained under a bias towards contractions.

Given the low quality of our parameter estimates, any conclusions drawn from this analysis are extremely tentative. Notwithstanding this, we suspect that our data (particularly for motif lengths 5 and 6) is overly polluted with non-microsatellite sequences. We suspect that this has led to the preference for very small values of c in our attempt at parameter estimation. Further work will be needed to find a reliable optimisation routine in order to get accurate parameter estimates. Once such a routine is determined, refitting the model to this data will allow us to investigate this further.

Our model does not rely on the equilibrium assumption, which is often dubious in the

context of biological evolution. The inclusion of point mutation provides an appropriate relative clock with which to fit a transient distribution to data. Further, the model accounts for the theorised life-cycle [18; 85; 96; 114] of microsatellite sequences from birth to eventual death, rather than assuming omnipresent sequences which change only in their repeat number. We believe that this makes for a meaningful progression of microsatellite models in terms of biological realism. However, finding maximum likelihood estimates with the model is difficult. As of now we do not have a working routine to achieve this. If after careful analysis of the optimisation problem we can find an appropriate routine, then data-driven analysis with this model can proceed. Otherwise, some theoretical analysis is still possible. If parameters of the model can be inferred from empirical work, then predictions about the life-cycle of microsatellites can still be made.

Chapter 6

Conclusions

In this thesis, we have developed and analysed models for duplicate gene evolution, and models for microsatellite evolution, which are two key areas of interest in evolutionary biology. In order to make accurate inferences in the context of evolutionary biology, particularly in the highly model-based applications of phylogenetics, realistic models for the evolution of sequences are required. The models constructed here are based on the biological mechanisms driving the evolution of gene duplicates and microsatellites, with each model representing a new mathematical perspective on the biological processes. We have analysed these models and fit them to data, arriving at new biological insights of the processes we model, as well as new mathematical results. We detail these below.

We have constructed an absorbing continuous-time Markov model for the evolution of a pair of gene duplicates evolving under the biological process of regulatory subfunctionalization. The model is based on the mechanics of the duplication-degeneration-complementation process described by Force [42]. We have introduced a modified-cause-specific hazard rate in order to analyse the hazard rate in the context where the process may have become immune to failure at some point in the past. We have analysed this rate function in general, and in the specific context of duplicate gene evolution. We have extended the model for pairs of gene duplicates to a population of gene duplicates by modelling the birth of such pairs as a Poisson process, and derived the distribution of pairs of gene duplicates preserved in the genome at any time. We have fit this distribution to age data of preserved duplicates in four mammalian genomes.

We have extended this model to model the evolution of a pair of gene duplicates evolving under the combined processes of sub- and neofunctionalization. We have

applied the general results derived in the analysis of the subfunctionalization-only model to this model, and we have fit the model to the same dataset discussed in the preceding paragraph.

Through the analysis of our models and data for the evolution of gene duplicates, we have arrived at the conclusion that the pattern of gene duplicate preservation is more consistent with sub- than with neofunctionalization. This finding contradicts earlier model-based analysis of sub- and neofunctionalization [53], but supports the hypothesis that subfunctionalization provides a protective mechanism allowing for subsequent neofunctionalization [42; 97]. Future work will investigate the dynamics of evolution of whole gene families using the model introduced in Section 4.2.4, using the approach employed in Section 3.8.

We have developed a model for the evolution of fixed-size families of gene duplicates, where multiple duplication events have taken place, which will be the subject of the future analysis mentioned above. We have applied a procedure for model development wherein an intuitive-to-develop, but intractable-to-analyse, model for a processes is mapped to a less intuitive, but more tractable model. By doing so, we are able to model all of the interactions between many gene duplicates without the need to explicitly consider the full set of possible transitions of the process envisioned as a Markov chain. In this way, we have provided a framework for modelling processes with simple underlying mechanisms but complex rules determining when those mechanisms are applied. We have outlined some computational considerations associated with this model, including an efficient framework for computing the generator of the process. We have further extended this model for the evolution of fixed-size gene families to dynamic-size gene families, using the same procedure.

We have constructed a level-dependent quasi-birth-and-death model for microsatellite evolution. The first discussion of this model was introduced during my honours [110], but is significantly refined here. We have first constructed a model where the number of repeat units, and number of repeat units which are not perfect matches to the motif are tracked as the level-variable and phase-variable respectively. In light of the consideration of whole-genome derived microsatellite data, we have redefined the model such that the phase-variable tracks the number of mismatches at the level of neucleotides, instead of the number of imperfect units. We have then considered two models, one in which the process is absorbed at some level-dependent cutoff in the number of mismatches, and one in which the process has no absorbing states. Our initial analysis indicated that treating the buildup of impurities as absorbing is the most realistic approach to modelling interrupted microsatellite sequences, and hence

we proceeded with this model.

In consideration of the absorbing model, we have discussed the possibility of fitting a limiting-conditional distribution to microsatellite data. Like the stationary distribution, which is very often fit to microsatellite data via likelihood-based methods, fitting the limiting conditional distribution requires the assumption that data is observed at equilibrium frequencies. We have considered an implementation of the return-map iterative scheme in which the ratio of means distribution is computed in place of the stationary distribution of the returned process. We have shown that this implementation is significantly more computationally efficient than the alternative, requiring only vector multiplication in place of Gaussian elimination. We have also shown that, under certain conditions, the scheme converges to the Yaglom limit associated with the scheme's starting vector, even for the case of a reducible transient set where the Yaglom limit is non-unique.

We have developed upon the analysis we introduced for the evolution of gene duplicates to derive a different distribution which is fit to microsatellite data. We have derived the transient distribution of the population of microsatellites under the assumption that the birth of microsatellites is a Poisson process. This provides a means to directly test the often-used assumption that the empirical data is observed at equilibrium frequencies, as a well as a means to fit to data regardless of whether it is or is not at equilibrium. We have fit this distribution to data using analogous likelihood-based methods to the stationary distribution, and we have used the rate of point mutations as a relative clock so that we do not need to explicitly consider time as a variable.

The data-driven analysis of microsatellite evolution was less successful. Ultimately, we were only able to conclude that the dataset which we analysed was too polluted with non-microsatellite sequence data to make meaningful inferences about the evolution of microsatellites. We have considered in some detail the difficulties associated with whole-genome derived data for impure microsatellites which led to the pollution of this data. Future work will fit the models for microsatellite evolution to a dataset with stricter purity requirements, which will be submitted for publication along with the results pertaining to microsatellite evolution discussed here.

All of the models we have constructed in this thesis operate at the level of individual sequences (or pairs/families of sequences), and are based on the mechanics of mutation which drive the evolution gene duplicates/microsatellites. Additionally, most of the models we have constructed are absorbing continuous time Markov chains. Usually, extension of individual-level models to the level of populations is achieved via the stationary distribution, which requires the assumption that empirical distributions

are observed at equilibrium frequencies (as well as models for which a meaningful stationary distribution exists). For each of the absorbing models, we have extended the individual-level model to the level of a population in a manner which does not rely on the widely-used assumption that empirical distributions are observed at equilibrium frequencies. This approach allows us to fit to data without the equilibrium assumption, as well providing a means to directly test the validity of such assumption. We have derived distributions for the population-level process in terms of the distribution over the state space of the individual level process, and in terms of the distribution of ages of surviving individuals. Both of these distributions can be calculated and fit to data with no more work than is required to calculate and fit the stationary distribution, with the exception that some particularly efficient methods exist for calculating stationary distributions.

We apply these results in the context of the evolution of gene duplicates and microsatellites, and the results were derived specifically for the analysis of these processes. Nonetheless, the approach in which we construct a model at the level of individuals, and then extend the model to a population, can be applied more broadly, and the results derived herein apply to models of this kind in general. It is likely that models for other evolutionary processes could follow a similar development, and these results could be applied directly. The results relevant to the modified-cause-specific hazard rate are applicable to processes which can either be killed, or become immune to subsequent killing; other applications could include any evolutionary process with a binary model for selection, or epidemiological models, where immunity is conferred to survivors of some disease. The results pertaining to the transient distribution of the population are applicable to population processes where new individuals are born with some fixed initial distribution. We have derived these results by employing results from the theory of absorbing continuous-time Markov processes, and the closely related phase-type distribution. The results presented here represent further corollaries to this theory. The scientific context serves as the impetus to consider a slightly different perspective on this theory, and hence as the impetus for the derivation of these results.

Future work will focus on further data-driven analysis with the models discussed here, particularly the model for microsatellite evolution and the model for the evolution of gene families of dynamic size. This will include application in a phylogenetic context, particularly the application of our models for gene duplication in ancestral state reconstruction. A more theoretical development we are pursuing is the strengthening of the result pertaining to the convergence of the ratio of means iteration to calculate quasi-stationary distributions. We did not end up applying quasi-stationary distribu-

tions here, but such distributions are of some practical and theoretical importance in probability modelling more generally. Most practical applications are restricted to the case where the quasi-stationary distribution is unique, but in the context where the initial distribution of the process is fixed, this need not be the case.

Appendix A

Tables of Microsatellite Fitting Results

Table A.1: Results for the full model fitting. Column labels are as follows: L is the Motif-length, u_0 and u_1 are the slippage constant and linear parameters respectively, b_0 and b_1 are the (logistic) constant and linear bias parameters respectively c is the exponential decay parameter associated with the build-up of mismatches, d is the rate of point mutation Lik is the likelihood associated with the fit, BIC is the BIC associated with the fit, $P(A_m)$ and $P(A_s)$ are the probabilities of absorption into the state associated a loss of repeatitive structure of the sequence due to point mutations, and contraction respectively, while $P(A_l)$ is absorption into the truncating state (and represents a sequence which at some point becomes longer than we allow for) and KL is the Kullback-Leibler Divergence score (lower score means a better goodness of fit). The genomes are, in order, Lizard, Lancelet, Nematode, Zebrafish, Fruitfly, Chicken, Human, Platypus and Penguin — each associated with $L = 1, ..., 6$.

Genome	L	u_0	u_1	b_0	b_1	c	d	Lik	BIC	$P(A_m)$	$P(A_s)$	$P(A_l)$	KL
Lizard	1	103.06	29.04	-2.34	5.72	9.55E-14	9.32	2.33E+06	4.65E+06	0.996	0.000	0.004	0.20

Table A.1: (continued)

Genome	L	u_0	u_1	b_0	b_1	c	d	Lik	BIC	$P(A_m)$	$P(A_s)$	$P(A_l)$	KL
Lizard	2	0.00	37.54	-7.15	4.90	2.55E-16	9.91	6.35E+05	1.27E+06	0.999	0.000	0.001	0.16
Lizard	3	1662.90	2588.54	-2.02	0.19	2.49E-13	362.82	6.74E+05	1.35E+06	0.536	0.449	0.015	0.37
Lizard	4	0.00	14.33	6.10	4.46	3.43E-11	9.40	2.25E+05	4.49E+05	1.000	0.000	0.000	0.11
Lizard	5	13.35	0.00	-5.89	3.15	1.60E+00	10.14	2.08E+05	4.16E+05	0.814	0.000	0.186	0.42
Lizard	6	0.00	0.30	5.96	7.96	1.37E+00	5.21	1.20E+05	2.39E+05	0.874	0.000	0.126	0.29
Lancelet	1	49.21	30.52	2.24	8.18	8.44E-12	9.49	9.17E+05	1.83E+06	0.997	0.000	0.003	0.17
Lancelet	2	0.51	244.65	-0.78	0.05	3.48E-15	11.66	2.05E+05	4.11E+05	0.224	0.775	0.000	0.13
Lancelet	3	0.00	14.39	-4.06	4.71	1.42E-11	9.53	9.46E+04	1.89E+05	1.000	0.000	0.000	0.09
Lancelet	4	0.81	53.85	-1.06	0.04	4.83E-12	3.28	9.86E+04	1.97E+05	0.195	0.804	0.001	0.10
Lancelet	5	0.00	0.22	8.75	8.56	2.16E-06	0.40	7.44E+04	1.49E+05	0.999	0.000	0.001	0.34
Lancelet	6	0.06	11.27	-2.47	0.16	2.09E-11	7.23	7.76E+04	1.55E+05	0.651	0.348	0.001	0.32
Nematode	1	0.00	32.17	-10.87	9.27	2.48E-01	9.95	3.88E+05	7.76E+05	0.998	0.000	0.002	0.34
Nematode	2	0.00	34.82	-4.93	0.66	6.24E-12	7.99	3.14E+04	6.30E+04	0.872	0.126	0.002	0.28
Nematode	3	8.89	257.83	-0.17	0.04	8.79E-05	52.99	2.19E+04	4.39E+04	0.587	0.412	0.001	0.17
Nematode	4	1.16	0.18	9.74	8.26	1.01E+00	1.20	6.84E+03	1.37E+04	0.993	0.000	0.007	0.18
Nematode	5	8.88	0.00	21.05	3.07	1.54E+00	10.36	8.49E+03	1.70E+04	0.934	0.000	0.066	0.33
Nematode	6	0.23	8.64	-2.31	0.08	1.02E-11	1.47	1.27E+04	2.55E+04	0.291	0.709	0.000	0.47
Zebrafish	1	46.73	14.33	-13.67	1.87	2.00E-11	3.42	4.45E+06	8.91E+06	0.810	0.189	0.002	0.15
Zebrafish	2	0.00	27.11	4.27	10.15	1.21E-10	3.36	1.51E+06	3.03E+06	1.000	0.000	0.000	0.38
Zebrafish	3	0.00	27.01	1.31	6.68	3.37E-15	10.00	3.47E+05	6.94E+05	1.000	0.000	0.000	0.13
Zebrafish	4	8.70E+03	4.02E+04	-1.24	0.05	3.63E-07	3155.42	3.72E+05	7.44E+05	0.238	0.762	0.001	0.11

Table A.1: (continued)

Genome	L	u_0	u_1	b_0	b_1	c	d	Lik	BIC	$P(A_m)$	$P(A_s)$	$P(A_l)$	KL
Zebrafish	5	0.00	1.73	-7.40	7.89	9.68E-01	9.73	2.33E+05	4.65E+05	1.000	0.000	0.000	0.30
Zebrafish	6	0.00	7.18	-8.37	6.76	2.73E-11	10.00	1.61E+05	3.23E+05	1.000	0.000	0.000	0.47
Fruitfly	1	0.00	50.37	5.86	9.52	4.39E-11	9.18	2.99E+05	5.99E+05	0.998	0.000	0.002	0.19
Fruitfly	2	0.00	45.08	8.67	9.85	7.24E-13	9.92	8.08E+04	1.62E+05	0.997	0.000	0.003	0.08
Fruitfly	3	0.05	13.10	-0.38	0.07	8.05E-11	2.46	5.27E+04	1.06E+05	0.622	0.376	0.002	0.18
Fruitfly	4	1.59	26.28	-2.83	0.25	1.99E-12	10.67	2.11E+04	4.23E+04	0.679	0.318	0.003	0.14
Fruitfly	5	0.00	0.08	30.40	1.22	1.25E+00	0.50	2.19E+04	4.39E+04	0.842	0.000	0.158	0.28
Fruitfly	6	0.00	18.05	-10.07	0.83	1.10E-09	12.02	3.21E+04	6.42E+04	0.738	0.260	0.001	0.36
Chicken	1	0.00	51.64	3.48	6.12	5.44E-11	9.53	5.26E+06	1.05E+07	0.999	0.000	0.001	0.14
Chicken	2	0.00	27.87	-4.65	0.58	1.63E-11	9.88	4.21E+05	8.42E+05	0.824	0.175	0.000	0.08
Chicken	3	0.00	10.59	7.02	3.20	4.54E-15	6.17	1.61E+05	3.21E+05	1.000	0.000	0.000	0.07
Chicken	4	0.00	15.48	-0.96	0.02	1.02E-03	0.34	1.31E+05	2.62E+05	0.075	0.924	0.001	0.04
Chicken	5	0.00	20.89	-10.36	1.00	4.61E-10	18.76	1.46E+05	2.92E+05	0.741	0.259	0.000	0.26
Chicken	6	0.00	0.57	12.40	5.75	1.40E+00	11.16	7.80E+04	1.56E+05	0.887	0.000	0.113	0.30
Human	1	93.03	18.64	-10.77	8.24	5.09E-13	3.72	9.67E+06	1.93E+07	0.998	0.000	0.002	0.22
Human	2	0.00	41.15	21.94	4.87	2.65E-15	8.97	1.12E+06	2.24E+06	1.000	0.000	0.000	0.17
Human	3	0.70	3320.29	-1.34	0.20	3.67E-04	1384.04	3.78E+05	7.57E+05	0.818	0.181	0.001	0.13
Human	4	0.00	1.67	-0.84	0.09	1.84E-03	0.45	3.64E+05	7.27E+05	0.619	0.378	0.002	0.07
Human	5	0.00	1.70	8.83	5.53	9.24E-01	8.77	3.13E+05	6.26E+05	0.997	0.000	0.003	0.32
Human	6	0.00	28.35	-4.84	0.35	5.65E-11	27.00	2.75E+05	5.50E+05	0.720	0.279	0.001	0.48
Platypus	1	83.90	30.42	1.52	3.78	9.43E-13	8.54	2.54E+06	5.08E+06	0.996	0.000	0.004	0.13

Table A.1: (continued)

Genome	L	u_0	u_1	b_0	b_1	c	d	Lik	BIC	$P(A_m)$	$P(A_s)$	$P(A_l)$	KL
Platypus	2	188.18	152.46	-2.05	0.23	1.53E-13	28.01	5.21E+05	1.04E+06	0.563	0.436	0.002	0.06
Platypus	3	0.00	17.72	-8.94	6.92	1.22E-01	9.84	7.98E+05	1.60E+06	0.998	0.000	0.002	0.06
Platypus	4	68.31	2.44	-1.45	0.04	4.31E-07	0.25	4.25E+05	8.50E+05	0.093	0.905	0.002	0.06
Platypus	5	1.34E+05	1.50E+07	-1.58E+06	1.58E+05	1.42E+00	1.55E+08	4.19E+05	8.38E+05	0.798	0.161	0.041	0.73
Platypus	6	1.60E+07	4.60E+08	-2.51E+08	9.95E+08	1.23E+00	6.00E+09	1.60E+05	3.21E+05	0.912	0.000	0.088	0.42
Penguin	1	0.00	40.07	3.80	8.62	2.36E-02	9.77	4.77E+06	9.53E+06	0.999	0.000	0.001	0.08
Penguin	2	0.00	23.32	0.63	3.63	9.06E-09	9.86	3.99E+05	7.98E+05	1.000	0.000	0.000	0.06
Penguin	3	0.00	15.62	-4.03	0.47	6.89E-12	10.00	1.43E+05	2.86E+05	0.908	0.091	0.000	0.07
Penguin	4	0.00	10.47	9.27	10.08	3.53E-14	10.00	1.05E+05	2.10E+05	1.000	0.000	0.000	0.06
Penguin	5	11.85	21.13	-2.12	0.08	1.43E-11	3.92	1.40E+05	2.81E+05	0.293	0.706	0.001	0.27
Penguin	6	0.00	2.81	10.21	7.09	2.25E-10	8.09	8.16E+04	1.63E+05	1.000	0.000	0.000	0.39

Table A.2: Results for the purity-independent model fitting. Column labels are as follows: L is the Motif-length, u_0 and u_1 are the slippage constant and linear parameters respectively, b_0 and b_1 are the (logistic) constant and linear bias parameters respectively, d is the rate of point mutation Lik is the likelihood associated with the fit, BIC is the BIC associated with the fit, ΔBIC is the difference between the BIC of the full (purity dependent) model and the BIC of the purity-independent model (negative means the purity-independent model is preferred) $P(A_m)$ and $P(A_s)$ are the probabilities of absorption into the state associated a loss of repetitive structure of the sequence due to point mutations, and contraction respectively, while $P(A_l)$ is absorption into the truncating state (and represents a sequence which at some point becomes longer than we allow for).

Genome	L	u_0	u_1	b_0	b_1	d	Lik	BIC	Δ BIC	$P(A_m)$	$P(A_s)$	$P(A_l)$
Lizard	1	100.00	0.00	-0.39	9.78	4.62	2.38E+06	4.75E+06	9.86E+04	0.99	0.01	0.00
Lizard	2	0.00	8.57	-9.20	9.94	6.08	7.31E+05	1.46E+06	1.93E+05	0.61	0.39	0.00
Lizard	3	0.00	24.65	1.08	9.14	9.74	7.59E+05	1.52E+06	1.69E+05	0.06	0.94	0.00
Lizard	4	0.00	1.76	4.57	8.26	5.15	2.47E+05	4.94E+05	4.46E+04	0.96	0.04	0.00
Lizard	5	21.12	0.00	-1.29	10.26	10.00	2.09E+05	4.18E+05	1.89E+03	0.99	0.01	0.00
Lizard	6	0.00	0.68	4.65	9.58	6.58	1.21E+05	2.42E+05	2.56E+03	1.00	0.00	0.00
Lancelet	1	95.68	0.00	1.97	3.46	4.75	9.47E+05	1.89E+06	6.04E+04	0.99	0.01	0.00
Lancelet	2	0.00	9.94	9.60	6.41	9.97	2.46E+05	4.92E+05	8.07E+04	0.87	0.13	0.00
Lancelet	3	0.22	0.27	86.02	164.37	0.83	9.58E+04	1.92E+05	2.22E+03	0.99	0.01	0.00
Lancelet	4	0.00	3.41	1.53	7.17	8.92	1.12E+05	2.25E+05	2.75E+04	0.96	0.04	0.00
Lancelet	5	7.27	0.71	-8.57	9.06	9.33	7.39E+04	1.48E+05	-9.88E+02	1.00	0.00	0.00
Lancelet	6	854.13	4.99E+07	-8.37E+10	6.98E+09	3.07E+08	7.96E+04	1.59E+05	3.94E+03	0.84	0.01	0.15
Nematode	1	97.06	0.00	-5.90	10.00	4.91	3.95E+05	7.89E+05	1.31E+04	0.99	0.01	0.00
Nematode	2	101.02	1.42	-3.98	9.75	8.87	3.26E+04	6.52E+04	2.32E+03	1.00	0.00	0.00
Nematode	3	19.36	2.32	8.99	9.16	9.14	2.17E+04	4.34E+04	-3.99E+02	0.99	0.01	0.00

Table A.2: (continued)

Genome	L	u_0	u_1	b_0	b_1	d	Lik	BIC	Δ BIC	$P(A_m)$	$P(A_s)$	$P(A_l)$
Nematode	4	4.95	1.03	-8.48	3.44	6.30	6.84E+03	1.37E+04	-7.67E+00	0.99	0.01	0.00
Nematode	5	7.63	0.65	5.84	8.51	9.87	8.56E+03	1.71E+04	1.34E+02	0.99	0.01	0.00
Nematode	6	0.00	1.03	-0.99	9.95	7.56	1.29E+04	2.58E+04	3.26E+02	1.00	0.00	0.00
Zebrafish	1	75.76	0.03	0.85	7.64	2.94	4.67E+06	9.34E+06	4.32E+05	0.99	0.01	0.00
Zebrafish	2	0.00	34.64	10.20	5.74	9.30	1.84E+06	3.67E+06	6.46E+05	0.12	0.88	0.00
Zebrafish	3	0.00	4.12	-2.96	5.90	8.61	3.75E+05	7.50E+05	5.56E+04	0.96	0.04	0.00
Zebrafish	4	0.00	37.50	-0.67	0.03	20.40	4.02E+05	8.05E+05	6.06E+04	0.24	0.06	0.70
Zebrafish	5	0.00	1.65	10.16	9.86	10.00	2.33E+05	4.65E+05	1.40E+02	1.00	0.00	0.00
Zebrafish	6	0.00	0.09	-10.53	8.13	0.54	1.58E+05	3.15E+05	-7.42E+03	0.99	0.01	0.00
Fruitfly	1	98.83	1.18	-1.54	8.87	4.29	3.07E+05	6.15E+05	1.61E+04	0.99	0.01	0.00
Fruitfly	2	12.84	9.48	-8.19	6.28	9.95	8.31E+04	1.66E+05	4.55E+03	0.83	0.17	0.00
Fruitfly	3	0.00	3.88	5.84	2.88	8.24	5.29E+04	1.06E+05	2.08E+02	0.96	0.04	0.00
Fruitfly	4	0.00	2.85	-10.13	9.66	9.12	2.10E+04	4.21E+04	-2.13E+02	0.96	0.04	0.00
Fruitfly	5	0.00	1.88	-10.21	9.92	9.57	2.21E+04	4.42E+04	3.62E+02	0.98	0.02	0.00
Fruitfly	6	0.00	1.85	0.10	9.69	10.00	3.16E+04	6.32E+04	-9.88E+02	0.98	0.02	0.00
Chicken	1	0.00	36.51	6.41	2.01	9.59	5.39E+06	1.08E+07	2.61E+05	0.68	0.32	0.00
Chicken	2	44.23	43.49	-0.01	0.00	10.63	4.37E+05	8.73E+05	3.10E+04	0.29	0.01	0.70
Chicken	3	2.06	3.58	6.94	2.36	10.02	1.64E+05	3.27E+05	5.97E+03	0.98	0.02	0.00
Chicken	4	0.01	1.56	-0.16	0.00	1.14	1.35E+05	2.70E+05	8.33E+03	0.38	0.00	0.61
Chicken	5	0.00	0.87	9.97	5.21	5.04	1.50E+05	2.99E+05	7.65E+03	1.00	0.00	0.00
Chicken	6	0.00	2.90	-1.35	0.12	17.52	7.84E+04	1.57E+05	8.63E+02	0.81	0.01	0.18

Table A.2: (continued)

Genome	L	u_0	u_1	b_0	b_1	d	Lik	BIC	Δ BIC	$P(A_m)$	$P(A_s)$	$P(A_l)$
Human	1	99.59	13.14	-5.69	4.66	7.82	1.06E+07	2.12E+07	1.85E+06	0.86	0.14	0.00
Human	2	0.00	12.71	-0.49	9.02	10.00	1.27E+06	2.54E+06	3.03E+05	0.69	0.31	0.00
Human	3	0.00	3.95	-2.67	7.98	10.00	3.91E+05	7.81E+05	2.48E+04	0.97	0.03	0.00
Human	4	0.00	3.22	-6.37	7.83	9.46	3.84E+05	7.68E+05	4.13E+04	0.96	0.04	0.00
Human	5	0.00	1.54	-9.96	9.75	8.95	3.13E+05	6.26E+05	2.39E+02	0.99	0.01	0.00
Human	6	0.00	0.96	1.15	6.84	10.00	2.74E+05	5.48E+05	-1.65E+03	1.00	0.00	0.00
Platypus	1	89.20	0.00	-10.09	10.05	3.83	2.62E+06	5.25E+06	1.64E+05	0.99	0.01	0.00
Platypus	2	0.00	4.89	8.09	6.05	4.88	5.69E+05	1.14E+06	9.52E+04	0.85	0.15	0.00
Platypus	3	0.00	5.12	-1.16	7.71	9.48	8.15E+05	1.63E+06	3.27E+04	0.86	0.14	0.00
Platypus	4	0.00	2.54	9.42	10.08	9.21	4.47E+05	8.94E+05	4.36E+04	0.97	0.03	0.00
Platypus	5	0.00	0.73	10.33	9.75	9.95	4.32E+05	8.63E+05	2.51E+04	1.00	0.00	0.00
Platypus	6	0.09	0.08	19.24	-0.38	0.66	1.61E+05	3.22E+05	8.75E+02	0.99	0.01	0.00
Penguin	1	1.71E+09	1.30E+10	0.61	-0.03	1.29E+09	4.80E+06	9.61E+06	7.06E+04	0.48	0.00	0.51
Penguin	2	29.68	5.09	8.29	8.92	9.96	4.08E+05	8.16E+05	1.74E+04	0.97	0.03	0.00
Penguin	3	0.00	3.58	-5.57	4.77	9.98	1.49E+05	2.98E+05	1.17E+04	0.98	0.02	0.00
Penguin	4	0.00	1.58	-3.19	3.48	6.55	1.07E+05	2.15E+05	4.89E+03	0.99	0.01	0.00
Penguin	5	0.00	1.20	1.28	9.90	8.22	1.42E+05	2.84E+05	2.81E+03	1.00	0.00	0.00
Penguin	6	0.00	0.81	10.20	1.32	9.99	8.02E+04	1.60E+05	-2.85E+03	1.00	0.00	0.00

Table A.3: Results for the constant-bias fitting. Column labels are as follows: L is the Motif-length, u_0 and u_1 are the slippage constant and linear parameters respectively, c is the exponential decay parameter associated with the build-up of mismatches, d is the rate of point mutation, β is the probability that a slippage event leads to an expansion, Lik is the likelihood associated with the fit, BIC is the BIC associated with the fit, ΔBIC is the difference between the full model BIC and the BIC of this model (negative means this model is preferred) $P(A_m)$ and $P(A_s)$ are the probabilities of absorption into the state associated a loss of repeatitive structure of the sequence due to point mutations, and contraction respectively, while $P(A_l)$ is absorption into the truncating state (and represents a sequence which at some point becomes longer than we allow for).

Genome	L	u_0	u_1	c	d	β	Lik	BIC	ΔBIC	$P(A_m)$	$P(A_s)$	$P(A_l)$
Lizard	1	226.4	64.8	5.52E-05	14.7	0.84	2.33E+06	4.65E+06	-2.37E+01	0.84	0.16	0.00
Lizard	2	0.0	76.1	6.05E-15	20.1	1.00	6.35E+05	1.27E+06	-1.20E+01	1.00	0.00	0.00
Lizard	3	0.0	97.6	1.43E-15	20.0	1.00	6.80E+05	1.36E+06	1.15E+04	0.99	0.00	0.01
Lizard	4	0.0	20.5	5.22E-15	13.4	1.00	2.25E+05	4.49E+05	-1.11E+01	1.00	0.00	0.00
Lizard	5	29.6	0.0	1.46E+00	12.3	0.66	2.08E+05	4.16E+05	-5.59E+02	0.75	0.18	0.07
Lizard	6	0.0	13.0	1.67E-13	27.2	1.00	1.23E+05	2.47E+05	7.74E+03	1.00	0.00	0.00
Lancelet	1	99.5	62.3	3.14E-14	19.3	1.00	9.17E+05	1.83E+06	-1.25E+01	1.00	0.00	0.00
Lancelet	2	0.0	151.0	6.19E-16	29.6	1.00	2.16E+05	4.31E+05	2.02E+04	1.00	0.00	0.00
Lancelet	3	0.0	28.2	8.71E-13	18.7	1.00	9.46E+04	1.89E+05	-1.03E+01	1.00	0.00	0.00
Lancelet	4	0.0	45.5	4.21E-15	19.2	1.00	1.03E+05	2.05E+05	7.91E+03	1.00	0.00	0.00
Lancelet	5	1.2	0.0	1.33E+00	0.4	0.60	7.32E+04	1.46E+05	-2.36E+03	0.75	0.22	0.03
Lancelet	6	0.0	12.8	4.93E-15	22.1	1.00	7.86E+04	1.57E+05	1.94E+03	1.00	0.00	0.00
Nematode	1	0.0	61.6	2.48E-01	19.1	1.00	3.88E+05	7.76E+05	-1.16E+01	1.00	0.00	0.00
Nematode	2	0.0	72.1	1.90E-14	16.5	1.00	3.15E+04	6.30E+04	3.80E+01	1.00	0.00	0.00
Nematode	3	0.0	40.9	8.75E-15	20.0	1.00	2.20E+04	4.40E+04	1.17E+02	1.00	0.00	0.00

Table A.3: (continued)

Genome	L	u_0	u_1	c	d	β	Lik	BIC	ΔBIC	$P(A_m)$	$P(A_s)$	$P(A_l)$
Nematode	4	0.0	20.3	1.67E-15	19.9	1.00	6.94E+03	1.39E+04	1.97E+02	1.00	0.00	0.00
Nematode	5	21.2	0.0	1.52E+00	20.9	0.88	8.49E+03	1.70E+04	-1.01E+01	0.91	0.04	0.05
Nematode	6	0.0	15.2	2.09E-12	20.0	1.00	1.33E+04	2.66E+04	1.10E+03	1.00	0.00	0.00
Zebrafish	1	4.7	101.5	5.98E-14	20.0	1.00	4.46E+06	8.91E+06	3.90E+03	1.00	0.00	0.00
Zebrafish	2	0.0	127.8	9.13E-10	15.8	1.00	1.51E+06	3.03E+06	-1.26E+01	1.00	0.00	0.00
Zebrafish	3	0.0	54.8	4.73E-12	20.3	1.00	3.47E+05	6.94E+05	-1.14E+01	1.00	0.00	0.00
Zebrafish	4	0.0	49.9	6.41E-12	17.2	1.00	3.91E+05	7.82E+05	3.81E+04	1.00	0.00	0.00
Zebrafish	5	0.0	3.6	9.68E-01	20.4	1.00	2.33E+05	4.65E+05	-1.16E+01	1.00	0.00	0.00
Zebrafish	6	0.1	209.4	3.26E-05	289.8	0.99	1.61E+05	3.23E+05	-8.32E+00	1.00	0.00	0.00
Fruitfly	1	0.0	45.2	2.62E-13	8.2	1.00	2.99E+05	5.99E+05	-1.12E+01	1.00	0.00	0.00
Fruitfly	2	0.0	49.0	4.79E-16	10.8	1.00	8.08E+04	1.62E+05	-9.84E+00	1.00	0.00	0.00
Fruitfly	3	0.0	55.0	6.13E-15	18.6	1.00	5.31E+04	1.06E+05	6.18E+02	1.00	0.00	0.00
Fruitfly	4	0.0	1.5	1.11E+00	5.8	1.00	2.10E+04	4.20E+04	-3.70E+02	0.84	0.00	0.16
Fruitfly	5	0.0	10.5	6.90E-15	13.9	1.00	2.26E+04	4.53E+04	1.38E+03	1.00	0.00	0.00
Fruitfly	6	0.0	18.9	3.99E-13	18.3	1.00	3.24E+04	6.48E+04	6.19E+02	1.00	0.00	0.00
Chicken	1	0.1	10785.5	2.46E-04	0.2	0.46	5.15E+06	1.03E+07	-2.06E+05	0.00	1.00	0.00
Chicken	2	0.0	53.0	1.37E-14	18.8	1.00	4.22E+05	8.45E+05	2.85E+03	1.00	0.00	0.00
Chicken	3	0.0	34.3	3.55E-14	20.0	1.00	1.61E+05	3.21E+05	-1.08E+01	1.00	0.00	0.00
Chicken	4	0.0	1.0	2.11E-04	0.4	0.81	1.32E+05	2.64E+05	2.39E+03	0.87	0.13	0.00
Chicken	5	0.0	12.5	2.26E-14	18.9	1.00	1.48E+05	2.96E+05	4.48E+03	1.00	0.00	0.00
Chicken	6	0.0	7.3	8.01E-15	18.0	1.00	7.99E+04	1.60E+05	3.94E+03	1.00	0.00	0.00

Table A.3: (continued)

Genome	L	u_0	u_1	c	d	β	Lik	BIC	ΔBIC	$P(A_m)$	$P(A_s)$	$P(A_l)$
Human	1	159.1	31.9	2.86E-12	6.4	1.00	9.67E+06	1.93E+07	-1.46E+01	1.00	0.00	0.00
Human	2	0.0	89.9	8.44E-16	19.6	1.00	1.12E+06	2.24E+06	-1.25E+01	1.00	0.00	0.00
Human	3	0.0	28.2	3.04E-15	14.3	1.00	3.80E+05	7.59E+05	2.65E+03	1.00	0.00	0.00
Human	4	0.0	31.7	4.29E-12	18.0	1.00	3.67E+05	7.34E+05	6.77E+03	1.00	0.00	0.00
Human	5	0.0	11.8	2.71E-14	19.1	1.00	3.14E+05	6.29E+05	2.91E+03	1.00	0.00	0.00
Human	6	0.0	8.4	2.12E-10	18.7	1.00	2.79E+05	5.58E+05	7.45E+03	1.00	0.00	0.00
Platypus	1	106.3	41.7	9.04E-06	6.2	0.74	2.54E+06	5.08E+06	-8.79E+01	0.72	0.28	0.00
Platypus	2	0.0	76.5	1.34E-15	19.4	1.00	5.27E+05	1.05E+06	1.11E+04	1.00	0.00	0.00
Platypus	3	0.0	28.6	1.22E-01	15.9	1.00	7.98E+05	1.60E+06	-1.23E+01	1.00	0.00	0.00
Platypus	4	0.0	21.0	5.11E-15	20.1	1.00	4.31E+05	8.62E+05	1.21E+04	1.00	0.00	0.00
Platypus	5	5.6	0.0	2.10E+00	20.4	1.00	4.21E+05	8.43E+05	4.28E+03	0.96	0.00	0.04
Platypus	6	7.4	0.0	1.41E+00	9.7	1.00	1.60E+05	3.19E+05	-1.49E+03	0.90	0.00	0.10
Penguin	1	0.1	42.2	3.40E-02	0.3	0.46	4.73E+06	9.46E+06	-7.60E+04	0.06	0.94	0.00
Penguin	2	0.0	47.3	6.40E-15	20.0	1.00	3.99E+05	7.98E+05	-1.18E+01	1.00	0.00	0.00
Penguin	3	0.0	28.3	1.69E-14	18.1	1.00	1.43E+05	2.87E+05	4.52E+02	1.00	0.00	0.00
Penguin	4	0.0	20.9	7.32E-15	19.9	1.00	1.05E+05	2.10E+05	-1.05E+01	1.00	0.00	0.00
Penguin	5	0.0	9.4	1.37E-14	18.5	1.00	1.42E+05	2.84E+05	3.03E+03	1.00	0.00	0.00
Penguin	6	0.0	29.8	7.73E-12	84.2	0.99	8.16E+04	1.63E+05	-9.86E+00	1.00	0.00	0.00

Table A.4: Results for the no-bias fitting. Column labels are as follows: L is the Motif-length, u_0 and u_1 are the slippage constant and linear parameters respectively, c is the exponential decay parameter associated with the build-up of mismatches, d is the rate of point mutation, Lik is the likelihood associated with the fit, BIC is the BIC associated with the fit, ΔBIC is the difference between the full model BIC and the BIC of this model (negative means this model is preferred) $P(A_m)$ and $P(A_s)$ are the probabilities of absorption into the state associated a loss of repetitive structure of the sequence due to point mutations, and contraction respectively, while $P(A_l)$ is absorption into the truncating state (and represents a sequence which at some point becomes longer than we allow for).

Genome	L	u_0	u_1	c	d	Lik	BIC	Δ BIC	$P(A_m)$	$P(A_s)$	$P(A_l)$
Lizard	1	160.4	95.7	6.62E-16	2.66	2.33E+06	4.65E+06	-5.32E+02	0.20	0.80	0.00
Lizard	2	0.0	96.5	9.48E-15	2.82	6.35E+05	1.27E+06	1.49E+03	0.23	0.77	0.00
Lizard	3	0.0	21.3	2.79E-12	0.21	6.57E+05	1.31E+06	-3.40E+04	0.13	0.86	0.01
Lizard	4	0.0	13.1	2.17E-12	1.86	2.25E+05	4.50E+05	5.66E+02	0.45	0.55	0.00
Lizard	5	33.3	0.0	1.35E+00	8.67	2.08E+05	4.17E+05	4.60E+02	0.63	0.34	0.03
Lizard	6	0.0	1.3	1.25E+00	7.53	1.20E+05	2.40E+05	2.83E+02	0.73	0.26	0.02
Lancelet	1	0.0	91.6	1.21E-11	2.64	9.17E+05	1.83E+06	9.16E+01	0.22	0.78	0.00
Lancelet	2	0.0	27.0	2.29E-13	0.39	2.15E+05	4.31E+05	1.98E+04	0.17	0.83	0.00
Lancelet	3	0.0	67.6	9.12E-14	9.61	9.47E+04	1.89E+05	9.88E+01	0.44	0.56	0.00
Lancelet	4	0.0	98.5	2.92E-13	6.69	1.03E+05	2.06E+05	8.42E+03	0.34	0.66	0.00
Lancelet	5	13.9	0.0	1.37E+00	4.74	7.34E+04	1.47E+05	-2.02E+03	0.70	0.29	0.01
Lancelet	6	0.0	14.8	6.56E-12	8.84	7.86E+04	1.57E+05	2.05E+03	0.71	0.29	0.00
Nematode	1	0.0	100.0	9.10E-16	2.75	3.90E+05	7.79E+05	2.95E+03	0.22	0.78	0.00
Nematode	2	0.0	94.1	2.07E-13	2.11	3.15E+04	6.30E+04	5.82E+01	0.20	0.80	0.00
Nematode	3	0.0	100.1	6.50E-13	8.69	2.20E+04	4.40E+04	1.37E+02	0.36	0.64	0.00
Nematode	4	0.0	37.9	3.13E-13	10.00	6.95E+03	1.39E+04	1.97E+02	0.55	0.45	0.00

Table A.4: (continued)

Genome	L	u_0	u_1	c	d	Lik	BIC	Δ BIC	$P(A_m)$	$P(A_s)$	$P(A_l)$
Nematode	5	38.7	0.0	1.39E+00	16.30	8.53E+03	1.71E+04	6.72E+01	0.73	0.26	0.01
Nematode	6	0.0	24.3	7.99E-14	10.00	1.33E+04	2.66E+04	1.15E+03	0.65	0.35	0.00
Zebrafish	1	0.0	130.2	6.25E-12	2.12	4.46E+06	8.91E+06	6.94E+03	0.17	0.83	0.00
Zebrafish	2	0.0	53.2	1.21E-05	0.26	1.49E+06	2.98E+06	-4.71E+04	0.10	0.90	0.00
Zebrafish	3	0.0	100.0	3.13E-17	5.39	3.48E+05	6.95E+05	8.24E+02	0.30	0.70	0.00
Zebrafish	4	0.0	8.4	6.03E-11	0.30	3.86E+05	7.73E+05	2.89E+04	0.26	0.74	0.00
Zebrafish	5	0.0	4.0	1.05E+00	9.33	2.32E+05	4.64E+05	-1.24E+03	0.69	0.31	0.00
Zebrafish	6	0.0	22.5	8.56E-11	9.99	1.62E+05	3.23E+05	3.00E+02	0.66	0.34	0.00
Fruitfly	1	0.0	97.5	2.07E-13	1.39	3.00E+05	5.99E+05	4.18E+02	0.16	0.84	0.00
Fruitfly	2	0.0	98.0	1.13E-11	2.02	8.08E+04	1.62E+05	1.12E+02	0.19	0.81	0.00
Fruitfly	3	0.0	65.0	6.36E-11	2.99	5.31E+04	1.06E+05	7.29E+02	0.28	0.72	0.00
Fruitfly	4	0.0	76.4	5.30E-12	9.70	2.14E+04	4.28E+04	4.07E+02	0.43	0.57	0.00
Fruitfly	5	1.80E+06	1.40E+09	1.11E+00	2.90E+09	2.20E+04	4.40E+04	1.40E+02	0.64	0.35	0.01
Fruitfly	6	0.0	4.8	1.08E+00	11.07	3.13E+04	6.26E+04	-1.59E+03	0.58	0.40	0.02
Chicken	1	0.0	114.0	2.45E-01	2.30	5.22E+06	1.04E+07	-7.76E+04	0.18	0.82	0.00
Chicken	2	0.0	97.3	6.24E-13	4.79	4.23E+05	8.46E+05	3.69E+03	0.28	0.72	0.00
Chicken	3	0.0	85.4	1.02E-11	9.89	1.61E+05	3.21E+05	2.40E+02	0.41	0.59	0.00
Chicken	4	0.0	61.0	8.89E-12	9.96	1.32E+05	2.65E+05	3.19E+03	0.47	0.53	0.00
Chicken	5	0.0	3.3	1.00E+00	6.28	1.47E+05	2.95E+05	3.05E+03	0.66	0.34	0.00
Chicken	6	0.0	1.5	1.27E+00	9.89	7.80E+04	1.56E+05	9.69E+00	0.74	0.24	0.02
Human	1	191.6	29.2	4.76E-15	0.37	9.64E+06	1.93E+07	-5.26E+04	0.12	0.88	0.00

Table A.4: (continued)

Genome	L	u_0	u_1	c	d	Lik	BIC	Δ BIC	$P(A_m)$	$P(A_s)$	$P(A_l)$
Human	2	0.0	104.1	3.93E-13	2.14	1.12E+06	2.24E+06	2.44E+03	0.19	0.81	0.00
Human	3	0.0	100.0	1.16E-15	9.19	3.80E+05	7.60E+05	3.26E+03	0.37	0.63	0.00
Human	4	0.0	89.4	1.83E-13	10.01	3.67E+05	7.35E+05	7.54E+03	0.41	0.59	0.00
Human	5	0.0	18.4	3.03E-11	10.00	3.15E+05	6.29E+05	3.26E+03	0.69	0.31	0.00
Human	6	0.0	12.4	1.27E-09	10.30	2.79E+05	5.58E+05	7.76E+03	0.76	0.24	0.00
Platypus	1	104.1	81.2	2.97E-15	1.91	2.54E+06	5.08E+06	3.89E+02	0.19	0.81	0.00
Platypus	2	0.0	94.1	7.58E-12	2.56	5.27E+05	1.06E+06	1.23E+04	0.22	0.78	0.00
Platypus	3	6.13E+05	8.17E+08	1.98E-02	7.75E+07	7.99E+05	1.60E+06	1.29E+03	0.36	0.64	0.00
Platypus	4	0.0	37.9	2.66E-14	9.65	4.31E+05	8.62E+05	1.25E+04	0.55	0.45	0.00
Platypus	5	137.3	981.7	1.48E+00	1.04E+04	4.27E+05	8.54E+05	1.57E+04	0.85	0.14	0.01
Platypus	6	7.1	0.0	1.35E+00	3.55	1.61E+05	3.23E+05	2.03E+03	0.71	0.28	0.01
Penguin	1	0.0	131.2	1.27E-01	3.68	4.75E+06	9.50E+06	-3.00E+04	0.22	0.78	0.00
Penguin	2	0.0	100.0	2.17E-03	6.72	4.00E+05	7.99E+05	7.41E+02	0.32	0.68	0.00
Penguin	3	0.0	74.4	1.27E-14	10.00	1.43E+05	2.87E+05	6.79E+02	0.43	0.57	0.00
Penguin	4	0.0	33.8	1.46E-11	8.56	1.05E+05	2.10E+05	1.36E+02	0.54	0.46	0.00
Penguin	5	0.0	4.1	1.03E+00	10.30	1.41E+05	2.81E+05	2.27E+02	0.71	0.29	0.00
Penguin	6	0.0	0.4	1.63E-08	0.38	8.13E+04	1.63E+05	-7.31E+02	0.78	0.22	0.00

Bibliography

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] M. I. Arnone and E. H. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–1864, 1997.

[3] J. R. Artalejo and M. Lopez-Herrero. Quasi-stationary and ratio of expectations distributions: a comparative study. *Journal of Theoretical Biology*, 266(2):264–274, 2010.

[4] J. Aspi, E. Roininen, M. Ruokonen, I. Kojola, and C. Vilà. Genetic diversity, population structure, effective population size and demographic history of the finnish wolf population. *Molecular Ecology*, 15(6):1561–1576, 2006.

[5] H. Baumann and W. Sandmann. Numerical solution of level dependent quasi-birth-and-death processes. *Procedia Computer Science*, 1(1):1561–1569, 2010.

[6] N. Bean, L. Bright, G. Latouche, C. Pearce, P. Pollett, and P. G. Taylor. The quasi-stationary behavior of quasi-birth-and-death processes. *The Annals of Applied Probability*, 7(1):134–155, 1997.

[7] N. Bean, P. Pollett, and P. Taylor. Quasistationary distributions for level-dependent quasi-birth-and-death processes. *Stochastic Models*, 16(5):511–541, 2000.

[8] M. A. Beaumont. Detecting population expansion and decline using microsatellites. *Genetics*, 153(4):2013–2029, 1999.

[9] G. Bell and J. Jurka. The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single-step mutation process. *Journal of Molecular Evolution*, 44(4):414–421, 1997.

[10] G. Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27(2):573, 1999.

[11] A. Bhargava and F. Fuentes. Mutational dynamics of microsatellites. *Molecular Biotechnology*, 44(3):250–266, 2010.

[12] D. Bratton and J. Kennedy. Defining a standard for particle swarm optimization. In *Swarm Intelligence Symposium*, pages 120–127. IEEE, 2007.

[13] L. Bright and P. G. Taylor. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Stochastic Models*, 11(3):497–525, 1995.

[14] B. Brinkmann, M. Klintschar, F. Neuhuber, J. Hühne, and B. Rolf. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *The American Journal of Human Genetics*, 62(6):1408–1415, 1998.

[15] J. Brohede, C. Primmer, A. Møller, and H. Ellegren. Heterogeneity in the rate and pattern of germline mutation at individual microsatellite loci. *Nucleic Acids Research*, 30(9):1997–2003, 2002.

[16] A. Brüniche-Olsen, M. E. Jones, J. J. Austin, C. P. Burridge, and B. R. Holland. Extensive population decline in the Tasmanian devil predates European settlement and devil facial tumour disease. *Biology Letters*, 10(11):20140619, 2014.

[17] K. Burnham and D. Anderson. *Model selection and multi-model inference: a practical information-theoretic approach*. Springer, 2002.

[18] E. Buschiazzo and N. J. Gemmell. The rise, fall and renaissance of microsatellites in eukaryotic genomes. *BioEssays*, 28(10):1040–1050, 2006.

[19] E. Buschiazzo and N. J. Gemmell. Conservation of human microsatellites across 450 million years of evolution. *Genome Biology and Evolution*, 2:153–165, 2010.

[20] P. Calabrese and R. Durrett. Dinucleotide repeats in the *Drosophila* and human genomes have complex, length-dependent mutation processes. *Molecular Biology and Evolution*, 20(5):715–725, 2003.

[21] P. Calabrese and R. Sainudiin. Models of microsatellite evolution. In *Statistical methods in molecular evolution*, Statistics for Biology and Health, pages 289–305. Springer, 2005.

[22] P. J. Cameron. *Permutation groups*, volume 45 of *London Mathematical Society Student Texts*. Cambridge University Press, 1999.

[23] L. Comtet. *Advanced combinatorics: the art of finite and infinite expansions*. Springer Science & Business Media, 2012.

[24] G. Cooper, N. Burroughs, D. Rand, D. Rubinsztein, and W. Amos. Markov chain Monte Carlo analysis of human Y-chromosome microsatellites provides evidence of biased mutation. *Proceedings of the National Academy of Sciences*, 96(21):11916–11921, 1999.

[25] J.-M. Cornuet and G. Luikart. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, 144(4):2001–2014, 1996.

[26] J.-M. Cornuet, V. Ravigné, and A. Estoup. Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics*, 11(1):401, 2010.

[27] R. Cristescu, W. B. Sherwin, K. Handasyde, V. Cahill, and D. W. Cooper. Detecting bottlenecks using BOTTLENECK 1.2.02 in wild populations: the importance of the microsatellite structure. *Conservation Genetics*, 11(3):1043–1049, 2010.

[28] R. Crossman. *Limiting conditional distributions: imprecision and relation to the hazard rate*. PhD thesis, Durham University, 2009.

[29] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes*, volume II: General Theory and Structure of *Probability and Its Applications*. Springer Science & Business Media, 2007.

[30] J. N. Darroch and E. Seneta. On quasi-stationary distributions in absorbing discrete-time finite Markov chains. *Journal of Applied Probability*, 2(1):88–100, 1965.

[31] J. N. Darroch and E. Seneta. On quasi-stationary distributions in absorbing continuous-time finite Markov chains. *Journal of Applied Probability*, 4(1):192–196, 1967.

[32] J. P. Demuth, T. De Bie, J. E. Stajich, N. Cristianini, and M. W. Hahn. The evolution of mammalian gene families. *PloS one*, 1(1):e85, 2006.

[33] A. Di Rienzo, A. Peterson, J. Garza, A. Valdes, M. Slatkin, and N. Freimer. Mutational processes of simple-sequence repeat loci in human populations. *Proceedings of the National Academy of Sciences*, 91(8):3166–3170, 1994.

[34] M. Dupuy, M. Stenersen, T. Egeland, and B. Olaisen. Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. *Human Mutation*, 23(2):117–124, 2004.

[35] J. Eisen. Mechanistic basis for microsatellite instability. In *Microsatellites: evolution and applications*, pages 34–48. Oxford University Press, 1999.

[36] H. Ellegren. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics*, 16(12):551–558, 2000.

[37] H. Ellegren. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, 5(6):435–445, 2004.

[38] A. Estoup, C. Tailliez, J. Cornuet, and M. Solignac. Size homoplasy and mutational processes of interrupted microsatellites in two bee species, *Apis mellifera* and *Bombus terrestris (Apidae). Molecular Biology and Evolution*, 12(6):1074–1084, 1995.

[39] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[40] P. Ferrari, H. Kesten, S. Martinez, and P. Picco. Existence of quasi-stationary distributions. A renewal dynamical approach. *The Annals of Probability*, 23(2):501–521, 1995.

[41] J. W. Fondon III, A. Martin, S. Richards, R. A. Gibbs, and D. Mittelman. Analysis of microsatellite variation in *Drosophila melanogaster* with population-scale genome sequencing. *PLOS ONE*, 7(3):e33036, 2012.

[42] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y.-L. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, 1999.

[43] M. Gardner, C. Bull, S. Cooper, and G. Duffield. Microsatellite mutations in litters of the Australian lizard *Egernia stokesii. Journal of Evolutionary Biology*, 13(3):551–560, 2000.

[44] G. C. Garriga, E. Junttila, and H. Mannila. Banded structure in binary matrices. *Knowledge and Information Systems*, 28(1):197–226, 2011.

[45] J. Garza, M. Slatkin, and N. Freimer. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Molecular Biology and Evolution*, 12(4):594–603, 1995.

[46] D. Goldstein, A. Linares, L. Cavalli-Sforza, and M. Feldman. An evaluation of genetic distances for use with microsatellite loci. *Genetics*, 139(1):463–471, 1995.

[47] B. Harr and C. Schlötterer. Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. *Genetics*, 155(3):1213–1220, 2000.

[48] B. Harr, B. Zangerl, and C. Schlötterer. Removal of microsatellite interruptions by DNA replication slippage: phylogenetic evidence from *Drosophila*. *Molecular Biology and Evolution*, 17(7):1001–1009, 2000.

[49] S. Hautphenne, M. Massaro, and P. Taylor. How old is this bird? The age distribution under some phase sampling schemes. *Journal of Mathematical Biology*, 75:1319–1347, 2017.

[50] X. He and J. Zhang. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, 169(2):1157–1164, 2005.

[51] Q. Huang, F. Xu, H. Shen, H. Deng, Y. Liu, Y. Liu, J. Li, R. Recker, and H. Deng. Mutation patterns at dinucleotide microsatellite loci in humans. *The American Journal of Human Genetics*, 70(3):625–634, 2002.

[52] D. Hudson. Interval estimation from the likelihood function. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(2):256–262, 1971.

[53] T. Hughes and D. A. Liberles. The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo-than subfunctionalisation. *Journal of Molecular Evolution*, 65(5):574–588, 2007.

[54] T. Hughes and D. A. Liberles. Whole-genome duplications in the ancestral vertebrate are detectable in the distribution of gene family sizes of tetrapod species. *Journal of Molecular Evolution*, 67(4):343–357, 2008.

[55] H. Innan and F. Kondrashov. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2):97–108, 2010.

[56] C. Jacq, J. Miller, and G. Brownlee. A pseudogene structure in 5S DNA of Xenopus laevis. *Cell*, 12(1):109–120, 1977.

[57] P. Jarne and P. Lagoda. Microsatellites, from molecules to populations and back. *Trends in Ecology & Evolution*, 11(10):424–429, 1996.

[58] A. Jones, G. Rosenqvist, A. Berglund, and J. Avise. Clustered microsatellite mutations in the pipefish *Syngnathus typhle*. *Genetics*, 152(3):1057–1063, 1999.

[59] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In *Mammalian protein metabolism*, volume 3, pages 21–132. Academic Press, 1969.

[60] M. Kayser, L. Roewer, M. Hedman, L. Henke, J. Henke, S. Brauer, C. Krüger, M. Krawczak, M. Nagy, T. Dobosz, R. Szibor, P. de Knijff, M. Stoneking, and A. Sajantila. Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *The American Journal of Human Genetics*, 66(5):1580–1588, 2000.

[61] Y. D. Kelkar, K. A. Eckert, F. Chiaromonte, and K. D. Makova. A matter of life or death: how microsatellites emerge in and vanish from the human genome. *Genome Research*, 21(12):2038–2048, 2011.

[62] Y. D. Kelkar, N. Strubczewski, S. E. Hile, F. Chiaromonte, K. A. Eckert, and K. D. Makova. What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. *Genome Biology and Evolution*, 2:620–635, 2010.

[63] M. Kijima. Quasi-stationary distributions of single-server phase-type queues. *Mathematics of Operations Research*, 18(2):423–437, 1993.

[64] M. Kimura and J. F. Crow. The number of alleles that can be maintained in a finite population. *Genetics*, 49(4):725–738, 1964.

[65] C. V. Kirchhamer, C.-H. Yuh, and E. H. Davidson. Modular cis-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo, and additional examples. *Proceedings of the National Academy of Sciences*, 93(18):9322–9328, 1996.

[66] A. Konrad, A. I. Teufel, J. A. Grahnen, and D. A. Liberles. Toward a general model for the evolutionary dynamics of gene duplicates. *Genome Biology and Evolution*, 3:1197–1209, 2011.

[67] S. Kruglyak, R. Durrett, M. Schug, and C. Aquadro. Distribution and abundance of microsatellites in the yeast genome can be explained by a balance

between slippage events and point mutations. *Molecular Biology and Evolution*, 17(8):1210–1219, 2000.

[68] S. Kruglyak, R. T. Durrett, M. Schug, and C. Aquadro. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences*, 95(18):10774–10778, 1998.

[69] J. Kuha. AIC and BIC comparisons of assumptions and performance. *Sociological Methods & Research*, 33(2):188–229, 2004.

[70] V. G. Kulkarni. *Modeling and analysis of stochastic systems*. CRC Press, 1995.

[71] Y. Lai and F. Sun. The relationship between microsatellite slippage mutation rate and the number of repeat units. *Molecular Biology and Evolution*, 20(12):2123–2131, 2003.

[72] G. Latouche and V. Ramaswami. *Introduction to matrix analytic methods in stochastic modeling*. ASA-SIAM Series on Statistics and Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1999.

[73] S. Leclercq, E. Rivals, and P. Jarne. DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. *Genome Biology and Evolution*, 2:325–335, 2010.

[74] M. Legendre, N. Pochet, T. Pak, and K. J. Verstrepen. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Research*, 17(12):1787–1796, 2007.

[75] A. Leopoldino and S. Pena. The mutational spectrum of human autosomal tetranucleotide microsatellites. *Human Mutation*, 21(1):71–79, 2003.

[76] R. A. Leslie. How not to repeatedly differentiate a reciprocal. *The American Mathematical Monthly*, 98(8):732–735, 1991.

[77] G. Levinson and G. Gutman. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, 4(3):203–221, 1987.

[78] D. A. Liberles, G. Kolesov, and K. Dittmar. Understanding gene duplication through biochemistry and population genetics. In *Evolution after gene duplication*, pages 1–21. Wiley-Blackwell, 2010.

[79] D. A. Liberles, A. I. Teufel, L. Liu, and T. Stadler. On the need for mechanistic models in computational genomics and metagenomics. *Genome Biology and Evolution*, 5(10):2008–2018, 2013.

[80] D. A. Liberles, M. D. Tisdell, and J. A. Grahnen. Binding constraints on the evolution of enzymes and signalling proteins: the important role of negative pleiotropy. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1714):1930–1935, 2011.

[81] M. Lynch and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, 2000.

[82] M. Lynch and A. Force. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459–473, 2000.

[83] M. Lynch, M. O'Hely, B. Walsh, and A. Force. The probability of preservation of a newly arisen gene duplicate. *Genetics*, 159(4):1789–1804, 2001.

[84] F. J. Mendoza-Torres. A note on two classical theorems of the Fourier transform for bounded variation functions. *Communications in Mathematics and Applications*, 7(2):73–80, 2016.

[85] W. Missier, S.-H. Li, and C.-B. Stewart. The birth of microsatellites. *Nature*, 381(6582):483, 1996.

[86] R. T. Nelson and R. Shoemaker. Identification and analysis of gene families from the duplicated genome of soybean using est sequences. *BMC genomics*, 7(1):204, 2006.

[87] M. F. Neuts. *Matrix-geometric solutions in stochastic models: an algorithmic approach*, volume 2 of *Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins University Press, 1981.

[88] S. Ohno. The enormous diversity in genome sizes of fish as a reflection of nature's extensive experiments with gene duplication. *Transactions of the American Fisheries Society*, 99(1):120–130, 1970.

[89] T. Ohta and M. Kimura. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research*, 22(2):201–204, 9 1973.

[90] P. J. Olver. *Applications of Lie groups to differential equations*, volume 107 of *Graduate Texts in Mathematics*. Springer Science & Business Media, 2000.

[91] O. Paun and E. Hörandl. Evolution of hypervariable microsatellites in apomictic polyploid lineages of *Ranunculus carpaticola*: directional bias at dinucleotide loci. *Genetics*, 174(1):387–398, 2006.

[92] C. E. Pearson, K. N. Edamura, and J. D. Cleary. Repeat instability: mechanisms of dynamic mutations. *Nature Reviews Genetics*, 6(10):729–742, 2005.

[93] T. Phung-Duc, H. Masuyama, S. Kasahara, and Y. Takahashi. A simple algorithm for the rate matrices of level-dependent QBD processes. In *Proceedings of the 5th International Conference on Queueing Theory and Network Applications*, pages 46–52. ACM, 2010.

[94] S. Piry, G. Luikart, and J. Cornuet. BOTTLENECK: a computer program for detecting recent reductions in the effective population size using allele frequency data. *Journal of Heredity*, 90(4):502–503, 1999.

[95] C. Primmer, N. Saino, A. Møller, and H. Ellegren. Unraveling the processes of microsatellite evolution through analysis of germ line mutations in barn swallows *Hirundo rustica*. *Molecular Biology and Evolution*, 15(8):1047–1054, 1998.

[96] C. R. Primmer and H. Ellegren. Patterns of molecular evolution in avian microsatellites. *Molecular Biology and Evolution*, 15(8):997–1008, 1998.

[97] S. Rastogi and D. A. Liberles. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evolutionary Biology*, 5(1):28, 2005.

[98] V. Rohatgi and E. Saleh. *An introduction to probability and statistics*, volume 910 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, 2011.

[99] O. Rose and D. Falush. A threshold size for microsatellite expansion. *Molecular Biology and Evolution*, 15(5):613–615, 1998.

[100] S. M. Ross. *Introduction to probability models*. Academic Press, 2014.

[101] R. Sainudiin, R. Durrett, C. Aquadro, and R. Nielsen. Microsatellite mutation models insights from a comparison of humans and chimpanzees. *Genetics*, 168(1):383–395, 2004.

[102] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3):441–446, 1975.

[103] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[104] F. A. Sefideh, M. J. Moon, S. Yun, S. I. Hong, J.-I. Hwang, and J. Y. Seong. Local duplication of gonadotropin-releasing hormone (GnRH) receptor before two rounds of whole genome duplication and origin of the mammalian GnRH receptor. *PLOS ONE*, 9(2):e87901, 2014.

[105] K. A. Selkoe and R. J. Toonen. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters*, 9(5):615–629, 2006.

[106] R. Sibly, J. Whittaker, and M. Talbot. A maximum-likelihood approach to fitting equilibrium models of microsatellite evolution. *Molecular Biology and Evolution*, 18(3):413–417, 2001.

[107] N. J. A. Sloane and S. Plouffe. *The encyclopedia of integer sequences*. Academic Press, 1995.

[108] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

[109] T. L. Stark, D. A. Liberles, B. R. Holland, and M. M. O'Reilly. Analysis of a mechanistic Markov model for gene duplicates evolving under subfunctionalization. *BMC Evolutionary Biology*, 17(1):38, 2017.

[110] T. L. Stark, M. M. O'Reilly, and B. R. Holland. Markov models for microsatellite mutation. Unpublished honours thesis, University of Tasmania, 2013.

[111] D. A. Steane, N. Conod, R. C. Jones, R. E. Vaillancourt, and B. M. Potts. A comparative analysis of population structure of a forest tree, *Eucalyptus globulus* (Myrtaceae), using microsatellite markers and quantitative traits. *Tree Genetics & Genomes*, 2(1):30–38, 2006.

[112] G. W. Stewart. A Krylov–Schur algorithm for large eigenproblems. *SIAM Journal on Matrix Analysis and Applications*, 23(3):601–614, 2002.

[113] J. F. Storz and M. A. Beaumont. Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution*, 56(1):154–166, 2002.

[114] S. Subramanian, R. K. Mishra, and L. Singh. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biology*, 4(2):R13, 2003.

[115] H. M. Taylor and S. Karlin. *An introduction to stochastic modeling*. Academic Press, 2014.

[116] A. I. Teufel, L. Liu, and D. A. Liberles. Models for gene duplication when dosage balance works as a transition state to subsequent neo-or sub-functionalization. *BMC Evolutionary Biology*, 16(1):45, 2016.

[117] A. I. Teufel, J. Zhao, M. O'Reilly, L. Liu, and D. A. Liberles. On mechanistic modeling of gene content evolution: birth-death models and mechanisms of gene birth and gene retention. *Computation*, 2(3):112–130, 2014.

[118] G. Tóth, Z. Gáspári, and J. Jurka. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research*, 10(7):967–981, 2000.

[119] C. Twerdi, J. Boyer, and R. Farber. Relative rates of insertion and deletion mutations in a microsatellite sequence in cultured cells. *Proceedings of the National Academy of Sciences*, 96(6):2875–2879, 1999.

[120] V. N. Uversky. Unusual biophysics of intrinsically disordered proteins. *Biochimica et Biophysica Acta (BBA) — Proteins and Proteomics*, 1834(5):932–951, 2013.

[121] E. A. van Doorn and P. K. Pollett. Quasi-stationary distributions for discrete-state models. *European Journal of Operational Research*, 230(1):1–14, 2013.

[122] D. Vere-Jones. Some limit theorems for evanescent processes. *Australian & New Zealand Journal of Statistics*, 11(2):67–78, 1969.

[123] J. Walsh. Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. *Genetics*, 115(3):553–567, 1987.

[124] J. Watkins. Microsatellite evolution: Markov transition functions for a suite of models. *Theoretical Population Biology*, 71(2):147–159, 2007.

[125] J. L. Weber. Informativeness of human $(\text{dC-dA})_n \cdot (\text{dG-dT})_n$ polymorphisms. *Genomics*, 7(4):524–530, 1990.

[126] J. Whittaker, R. Harbord, N. Boxall, I. Mackay, G. Dawson, and R. Sibly. Likelihood-based estimation of microsatellite mutation rates. *Genetics*, 164(2):781–787, 2003.

[127] M. Wierdl, M. Dominska, and T. Petes. Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics*, 146(3):769–779, 1997.

[128] T. Willems, M. Gymrek, G. Highnam, D. Mittelman, Y. Erlich, and 1000 Genomes Project Consortium. The landscape of human STR variation. *Genome Research*, 24(11):1894–1904, 2014.

[129] L. Wissler, L. Godmann, and E. Bornberg-Bauer. Evolutionary dynamics of simple sequence repeats across long evolutionary time scale in genus *Drosophila*. *Trends in Evolutionary Biology*, 4(1):e7, 2012.

[130] C. Wu and A. Drummond. Joint inference of microsatellite mutation models, population history and genealogies using transdimensional Markov chain Monte Carlo. *Genetics*, 188(1):151–164, 2011.

[131] X. Xu, M. Peng, Z. Fang, and X. Xu. The direction of microsatellite mutations is dependent upon allele length. *Nature Genetics*, 24(4):396–399, 2000.

[132] A. Yates, W. Akanni, M. R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, S. Keenan, I. Lavidas, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, M. Nuhn, A. Parker, M. Patricio, M. Pignatelli, M. Rahtz, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, E. Birney, J. Harrow, M. Muffato, E. Perry, M. Ruffier, G. Spudich, S. J. Trevanion, F. Cunningham, B. L. Aken, D. R. Zerbino, and P. Flicek. Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710–D716, 2016.

[133] L. Zane, L. Bargelloni, and T. Patarnello. Strategies for microsatellite isolation: a review. *Molecular Ecology*, 11(1):1–16, 2002.

[134] M. Živković. Massive computation as a problem solving tool. In *Proceedings of the 10th Congress of Yugoslav Mathematicians*, pages 113–128. University of Belgrade, Faculty of Mathematics, 2001.

[135] M. Živković. Classification of small (0,1) matrices. *Linear Algebra and its Applications*, 414(1):310–346, 2006.